University of Warwick

Doctoral Thesis

The Value of Magnetic Resonance Imaging in the Assessment of Degenerative Lumbar Spinal Stenosis

Author:
Dr Richard Gagen

Supervisors:
Prof. Charles Hutchinson
Dr Caron Parsons

A thesis submitted in fulfilment of the requirements for the degree of doctor of philosophy

in the

Populations, Evidence and Technologies Group
Warwick Medical School

Volume 1 of 1
9th July 2021

Contents

List of Figures

List of Tables

Acknowledgements

I am most grateful to the following people, without whose assistance this thesis would not have been possible.

To Professor Charles Hutchinson and Dr Caron Parsons, who have supervised my research and clinical activities for the past four years and provided a great deal of encouragement and support. To Prof. Sally Lamb, Dr Esther Williamson, Mrs Angela Garret, and the large number of other individuals whose work on the BOOST randomised controlled trial provided so much of the data analysed within this thesis. To Dr Nick Parsons, Dr Sian Taylor-Phillips, Mrs Samantha Johnson and Mr Robert Sneeth who provided helpful advice on statistics, systematic review design, literature search strategy and the surgical perspective on lumbar stenosis. To Dr Terry Jones, Dr Carmen Dragos and Dr Ruth Hartley for their advice, work as MRI assessors and help with study blinding. To Dr Emily Evans, Dr Nageena Suleman and Dr Samantha Low for their support and teaching, given while putting up with my endless complaints about the writing-up processes. To Dr Mark Blakeman and Dr Alison Page, who supported my application for time out of clinical training to conduct this research. To the radiographers, sonographers, radiology registrars and consultants of the University Hospitals Coventry and Warwickshire and Birmingham Children's Hospital for their ongoing support. To Dr Mike Juniper, Mr Ashley Walters and Dr Sam Ivell alongside CornyEars, ZeroZipZilch and Brittlestar for kindly distracting me from work and preventing my slow slide into insanity. To the large number of patients without whose kind participation this research would not have been possible.

I would also like to thank Mr John E. Nixon (1987) and Mr R. W. Porter (2001), whose theses helped illuminate the history of research relating to lumbar spinal stenosis and also gave many insights into the pathophysiology of the condition. Similarly, the books *"Clinical and Radiological Anatomy of the Lumbar Spine"* by Professor Bogduk (2012) and *"Lumbar Spinal Imaging in Radicular Pain and Related Conditions"* by Professor Wilmink (2010b) have been highly influential on my thinking and are cited extensively in the introduction. Professor Ian H. Witten's book (2017) and online courses also provided a thorough but highly approachable introduction to the subject of machine learning, which I relied upon extensively in Chapter 8.

All statistical analyses presented in this thesis were performed in the *R environment for statistical computing* (R Core Team 2020), particularly using the *tidyverse* (Wickham 2017), *tidymodels* (Kuhn et al. 2020), *metaphore* (Viechtbauer 2010), *pROC* (Robin et al. 2011) and *caret* (Kuhn 2020) packages. In addition, several scripts were written for processing digital imaging and communications in medicine (DICOM) format images and modelling nerve root compression in *Python*, in particular using the *pydicom* and *pandas* packages. DICOM images were viewed in *Horos*, an open source DICOM viewer (Project 2019). I would like to thank the numerous contributors to these open source projects, who give their time free of charge to develop these fantastic tools.

This work was in part funded by National Institute of Health Research (NIHR) Program Grants for Applied Research (PTC-RP-PG-0213-20002).

Declaration

I, RICHARD GAGEN, declare that this thesis titled, 'The Value of Magnetic Resonance Imaging in the Assessment of Degenerative Lumbar Spinal Stenosis' and the work presented in it are my own.

I confirm that:

- This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.
- The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:
  - The protocol for the BOOST randomised controlled trial was designed by the BOOST trial team with input from Prof. Hutchinson and myself relating to the imaging aspects of the trial design.
  - The non-imaging clinical data for BOOST participants was collected by members of the BOOST team at each trial site.
  - The MRI studies were performed by radiographers at each trial site.
  - That two additional MRI assessors contributed to MRI data extraction as part of assessment of the reliability of this process.
- Where I have consulted or quoted the published work of others this is always clearly attributed.
- Parts of this thesis have been published by the author:
  - R. M. Gagen et al. (2017a). *MRI in degenerative lumbar spinal stenosis: is there a relationship between the dural-sac cross-sectional area and patient pain, disability and quality of life? A systematic review*. PROSPERO 2017 CRD42017064865 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42017064865
  - R. M. Gagen et al. (2017b). *MRI in degenerative lumbar spinal stenosis: what MRI findings are associated with neurogenic claudication? A systematic review*. PROSPERO 2017 CRD42017064873 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42017064873
  - E. Williamson et al. (Oct. 2018). 'Better Outcomes for Older People with Spinal Trouble (BOOST) Trial: A Randomised Controlled Trial of a Combined Physical and Psychological Intervention for Older Adults with Neurogenic Claudication, a Protocol'. In: *BMJ Open* 8.10, e022205
  - Gagen, RM et al. (2020). 'Can MRI of the Lumbar Spine Identify Neurogenic Claudication Patients Reliably? - Comparison of Single Variable and Machine Learning Approaches.' In: *Proceedings of the International Society of Magnetic Resonance in Medicine*. ISMRM. Virtual

Signed: _____

Date: _____

UNIVERSITY OF WARWICK

DOCTORAL THESIS ABSTRACT

The Value of Magnetic Resonance Imaging in the Assessment of Degenerative Lumbar Spinal Stenosis

Richard Gagen

This thesis explores the role of magnetic resonance imaging (MRI) of the lumbar spine in patients with the main clinical feature of lumbar spinal stenosis (LSS): neurogenic claudication (NC). NC is thought to be caused by positional compression of the cauda equina in a spinal canal narrowed by degenerative change. MRI is the primary tool for demonstrating such degeneration but no universally accepted and evidence based imaging definition of LSS exists.

Systematic reviews of the literature are presented: the first finds the available studies comparing MRIs in NC patients to a control group have unsuitable methodologies to propose a definition of stenosis, largely due to use of imaging based inclusion criteria. The second finds the strength of relationship between canal size and symptom severity in LSS patients is inconsistent across different studies, but with most papers using surgical patient cohorts, likely to exclude those with minor symptoms.

A diagnostic cross-sectional study, including both community and secondary care based participants is described, comparing MRIs in participants with NC and a separately recruited control group. Unlike prior studies, NC patients are selected for inclusion based upon their clinical presentation alone. NC patients are found to have smaller canals than the control group, but measurements of canal narrowing or qualitative judgement of nerve root compression generally fail to accurately predict NC symptoms, and various methods of combining the measurements, including machine learning techniques, fail to improve diagnostic accuracy. No convincing relationship between symptom severity and canal size is identified.

A definition for radiological LSS is proposed for the central canal (grade C — Schizas et al. 2010) the lateral recess (grade 2 nerve root entrapment – Bartynski et al. 2003), and the neural exit foramen (neural exit foramen depth less than 4 mm) based upon the best performing measurements and other pragmatic considerations.

Abbreviations

6-MWT	6 minute walk test
AMT	abbreviated mental test
ANOVA	analysis of variance
APCR	antero-to-posterior canal ratio
APD	antero-to-posterior canal diameter
API	application programming interface
AP	anterior-to-posterior
AUC	area under the curve
BAI	Beck anxiety index
BDI	Beck depression inventory
BMI	body mass index
BOOST	better outcomes for older people with spinal troubles
CART	classification and regression tree
CD	compact disc
CI	confidence interval
COMI	core outcome measures index
CSA	cross-sectional area
CSF	cerebrospinal fluid
CT	computed tomography
DICOM	digital imaging and communications in medicine
DS-APD	dural-sac antero-to-posterior diameter
DS-CSA	dural-sac cross-sectional area
DS-TD	dural-sac transverse diameter
EQ-5D	EuroQol 5-dimensions
ETL	echo train length
FOV	field of view
GCP	good clinical practice
GP	general practitioner
ICC	intra-class correlation
ID	identification
IEP	image exchange portal
IFD	inter-facet canal diameter
ILD	interligamentous canal diameter
IPD	inter-pedicular canal diameter
JOA	Japanese Orthopaedic Association
LBPBI	low back pain bothersomeness index
LBP	lower back pain
LPBI	leg pain bothersomeness index
LRA	lateral recess angle
LRD	lateral recess depth
LRH	lateral recess height
LSS	lumbar spinal stenosis
ML	machine learning
MRI	magnetic resonance imaging
MWD	maximum walking distance
NASS	North American Spinal Society
NC	neurogenic claudication
NEX	number of excitations
NFD	neural exit foramen diameter
NFH	neural exit foramen height
NHS	National Health Service
NIHR	National Institute of Health Research
NIH	National Institute of Health
NNC	number of nerve roots compressed
NNMC	number of nerve roots multiply compressed
NRS	numeric rating scale
ODI	Oswestry disability index
OPAL	Oxford pain, activity and lifestyle cohort study
PACS	picture archiving and communication system
PCA	principle component analysis
PLL	posterior longitudinal ligament
PSQI	Pittsburgh sleep quality index
QQ	quantile-quantile (plot)
RCT	randomised controlled trial
ROC	receiver operator characteristic
ROSE	random over sampling examples
SBI	stenosis bothersomeness index
SD	standard deviation
SF-36	Short-Form 36 questionnaire
SOP	standard operating procedure
SPPB	short physical performance battery
SR	stenosis ratio
SSD	solid state drive
SSDC	sub-articular sagittal diameter of the canal
TE	time to echo
TFI	Tilburg frailty index
TLA	three letter acronym
TR	time to repeat
UHCW	University Hospital Coventry and Warwickshire
UK	United Kingdom of Great Britain and Northern Ireland
US	ultrasound
USA	United States of America
VAS	visual analogue scale
VF-CSA	vertebral foramen cross-sectional area
WD	walking distance
ZCQ NI	Zurich claudication questionnaire neuro-ischaemic domain
ZCQ PD	Zurich claudication questionnaire pain domain
ZCQ SS	Zurich claudication questionnaire symptom subscale
ZCQ	Zurich claudication questionnaire

Dedicated to my family…

A science which orders its thoughts too early is stifled

Jakob Bronowski, The Common Sense of Science, 1959

Chapter 1

Introduction

1.1 A note on terminology…

A prospective attempt at reducing confusion.

The terms syndrome (a set of medical signs and symptoms correlated with each other — OED-Online (2019c)) and disease (a particular abnormal condition negatively affecting the structure or function of an organism — OED-Online (2019a)) are often confused or used interchangeably. This occurs, understandably, due to the often close pathological link that ties a syndrome and the particular disease causing it together.

Similarly the terms LSS and neurogenic claudication (the hallmark symptom associated with narrowing of the lumbar canal) are often conflated. Despite this, the pathological link between neurogenic claudication and the anatomical narrowing ascribed to cause it is far from clearly understood. It is possible for patients with apparently severe canal narrowing to have no symptoms and, vice versa, patients with a classical syndromic presentation of LSS may have no detectable abnormality of the lumbar spinal canal (Andreisek et al. 2011).

In order to profitably discuss lumbar spinal stenosis it is necessary to separate its various parts and, as such, I adopt the following general definitions within this thesis:

- *Intermittent claudication* — buttock and/or leg pain induced by exercise and relieved by rest.
- *Vascular claudication* — intermittent claudication relating to exercise induced ischaemia of the lower limb secondary to vascular insufficiency.
- *Neurogenic claudication* — intermittent claudication symptoms thought to be related to exercise exacerbated pathology affecting the lumbar nerve roots.
- *Anatomical lumbar spinal stenosis* — abnormal narrowing of the lumbar spinal canal to the level where it potentially compresses or otherwise impairs the function of the canal contents. This term is used to refer to anatomical changes in isolation from the symptomatic status of the patient.
- *Symptomatic lumbar spinal stenosis* — abnormal narrowing of the canal where the patient is symptomatic and lumbar canal narrowing is the presumed cause of this.

Use of the many synonyms of the terms above that occur in the literature (e.g. pseudoclaudication, spondylotic caudal radiculopathy, etc.) are avoided.

In addition, when discussing lumbar spinal stenosis in the context of its aetiology the following terms may be used.

- *Degenerative lumbar spinal stenosis* — anatomical lumbar spinal stenosis caused by accumulated lumbar degenerative change. These changes include hypertrophy of the facets and ligamentum flavum, bulging of the disc and spondylolisthesis.
- *Congenital lumbar spinal stenosis* — anatomical lumbar spinal stenosis manifesting clinically during childhood and caused by abnormal foetal development, for instance congenital malformations of the spine.
- *Developmental lumbar spinal stenosis* — lumbar canal narrowing, genetic or environmental in origin, caused by reduced growth of the lumbar canal and only manifesting clinically when growth is complete and superimposed upon subsequent degenerative change.

The separate definitions of congenital and developmental LSS were established by Verbiest (1976). This thesis almost exclusively deals with degenerative LSS, primarily occurring in those over 50 years of age. Developmental narrowing of the canal is considered where it may act as a predisposing factor for degenerative symptomatic LSS. Congenital causes of canal stenosis may be used to illustrate potential pathophysiological concepts but otherwise are largely excluded from the subsequent work.

In general, unqualified use of the term lumbar spinal stenosis is avoided as I find different people mean different combinations of the above when using it.

1.2 A history of low back and leg pain…

A history of knowledge relating to the spinal canal and the first descriptions of lumbar spinal stenosis.

1.2.1 Early descriptions of the spinal canal

Some of the earliest surviving descriptions of the spinal canal occur in the writings of Galen (Viale 2004). Galen lived during the second century AD, six centuries after Hippocrates, and rose to fame after becoming the personal physician of the stoic philosopher Marcus Aurelius, last of the "five good emperors" of Rome (Machiavelli 1517). It is not clear how much of Galen's contribution to knowledge regarding the spine is original; he in part learnt his craft at the Alexandrian schools of anatomy and almost certainly was influenced by earlier authors (Viale 2004). Established by the Ptolemaic dynasty, the Alexandrian schools had allowed anatomical knowledge to flourish, but by the time of Galen human dissection was largely forbidden and and the schools long past their prime. Galen's practical anatomy was derived from studying the wounds of gladiators and the dissection of animals (Marketos et al. 1999).

Galen described the spine largely in teleological terms. For example, in *de usu partium corporis humani* (on the usefulness of the parts of the body) Galen discusses the balance between the number of vertebra and the strength and function of the spine—fewer vertebra he believes would favour a rigid and stronger spine, while more vertebra allow greater movement, vital for life, but also the source of the spines vulnerability (translation: Galen et al. 1968). He also points out a large number of vertebra allows the spine to bend in a circular rather than angular manner, avoiding injury to the spinal cord.

In *de ossibus ad tirones*—on bones for beginners (translation: Galen 2011b)—and *de nervorum dissectione*—on the dissection of nerves (translation: Galen 2011a)—Galen describes the anatomy of the canal and its contents. The canal is formed by the vertebral arch extending from the vertebral body anteriorly, the bony structures surrounding and protecting the spinal cord and acting as sites for articulation between vertebra to allow flexion. The cord itself is surrounded within the canal by an extension of the meninges (the dural sac). Each nerve root originating from the cord leaves the spine laterally through rounded foramina formed between each adjacent vertebra. The lateral position of the foramina, argues Galen, avoids an overly long course for the nerves, risking injury, while also not passing more directly through the vertebral bodies as that would jeopardise their strength. Galen understood that the cord constituted a continuation of the brain and that it was responsible in part for orchestrating movement of the trunk and limbs. Through experimental transections on animals he mapped the both segmental nature of the cord and the sensory and motor innervation of the nerve roots (Marketos et al. 1999).

Galen's descriptions dominated Western medical thinking until the sixteenth century when his work began to be challenged as the Church's authority waned and human dissection again began to be practised. Leonardo da'Vinci included in his anatomical studies depictions of the spine, spinal cord and nerve roots showing a clear progression in accuracy over time (Bowen et al. 2017). In 1543, Andreas Vesalius published *de humani corporis fabrica* (on the fabric of human body), based upon human dissection, and freed medicine from Galen's grasp by showing multiple errors in the masters work. Despite this *de fabrica* added little in terms of spinal anatomy to Galen's original descriptions (Marketos et al. 1999) and many parts of the discussion relating to the spine clearly echo the arguments on spinal anatomy and function from 1200 years earlier (Richardson et al. 1998).

1.2.2 Getting to cauda equina compression…

In 1804, Antoine Portal, French physician and future founder of the Académie National de Médecine, first suggested abnormal lower limb neurology could be caused by changes in the size of the spinal canal (Portal 1804). This suggestion however seems to have had relatively little effect on prevalent medical thinking of the time and did not inspire much further investigation. Instead the literature related to spinal causes of leg symptoms became dominated by exploring vascular aetiologies, likely secondary to Jean-Martin Charcot's highly influential description of intermittent claudication in 1859 in a patient with an aortic aneurysm (Nixon 1987). Charcot's patient complained of leg pain and dysfunction brought on by walking short distances and relieved by rest. This was attributed to ischaemia of the muscles and nerves of the legs brought on by exercise (Charcot 1859). Similarly, Bramwell (1886) reported cases in which vascular insufficiency related to aortic aneurysms appeared to cause neurological symptoms in the lower extremities. Unlike Charcot, Bramwell thought the symptoms could be generated by intermittent ischaemia of the lower spinal cord and nerve roots rather than through direct ischaemia of the nervous structures in the leg. Following on from this work, Dejerine in 1911 coined the term *intermittent claudication of the spinal cord*, a condition that he believed could mimic the vascular claudication described by Charcot. He described a case of a 37 year old women with a syphilitic vasculitis and paraparesis brought on by walking. This was accompanied by variable upper motor neuron signs and disturbance of micturition (Dejerine 1911).

Oppenhiem and Kause began to turn interest towards compressive pathologies affecting the cauda equina by publishing the first description of cauda equina syndrome, or as the authors put it— strangulation of the cauda equina (Oppenheim et al. 1909). Shortly after, C.A. Elsberg described a patient with cauda equina related symptoms who had a markedly

enlarged ligamentum flavum found on operation (Elsberg 1913). Removal of the hypertrophied ligament appeared to completely relieve the patients symptoms.

In 1934, Mixter and Barr published their seminal paper linking disc pathology to radicular pain. The paper presented a case series of patients with sciatica who, on operation, were found to have compression of individual spinal nerve roots caused by a tear of the disc annulus and subsequent extrusion of disc material into the spinal canal. Surgical removal of the extruded disc material provided symptomatic relief (Mixter et al. 2009). Sciatica, previously thought to be caused by a rheumatic condition resulting in inflammation of the nerve itself, was transformed to a potentially surgically treatable condition (Wilmink 2010d). While this paper referred purely to radicular pain rather than claudication, it firmly established nerve root compression by degenerative processes as a cause for radiating back pain.

H.A. Brown in 1938, referring back to the previously mentioned paper by C.A Elsberg, presented a case series of seven patients with ligamentum flavum thickening, and wrote of the difficulty and sometimes impossibility of distinguishing patients with prolapsed discs from those with ligamentum flavum thickening clinically (Brown 1938). With time, the importance of the postural dependency of the symptoms became more apparent, for instance Van Gelderen (1948) described two cases of patients again with symptomatic hypertrophy of the ligamentum flavum whose lower limb pain was exacerbated by walking but not by cycling (an exercise in which the lumbar spine remains flexed), sharply distinguishing this presentation from vascular claudication, where posture during exercise would not be expected to have an effect (Wilmink 2010c).

Verbiest (1954) is often credited with the first true description of symptomatic lumbar spinal stenosis (for example: Moses et al. 2015). He presented a case series of seven patients with bilateral radicular symptoms, tiredness, loss of power and anaesthesia of the legs that was exacerbated by standing or walking but absent in recumbency. Verbiest commented that the symptoms could easily be mistaken for vascular claudication but on investigation each showed a myelographic block caused by general narrowing of the spinal canal—a contrast agent injected into the lower dural-sac failed to pass beyond the area of canal narrowing. This block appeared to ease in several of the patients when the patient assumed a recumbent position. Verbiest believed the patients' posture and exercise related symptoms were associated with positional compression of the cauda equina at the narrowed canal segment, exacerbated by extension of the lumbar spine, and later coined the term *neurogenic intermittent claudication* to describe the syndromic presentation (Verbiest 1976). Similar to the previous case studies described, operative relief of the canal narrowing and consequently the cauda equina compression appeared to relieved the patients symptoms.

The discovery that Verbiest's patients had markedly reduced diameters of the bony canal, both measured intraoperatively and on lateral radiographs, led to the belief that the observed canal narrowing was caused by abnormally short pedicles, likely secondary to a defect of development (Verbiest 1976). These findings lead to a search for diagnostic measurements for stenosis on plain films and this search dominated the next two decades of research on LSS (Nixon 1987).

1.2.3 Early attempts to define LSS

Stenosis is defined by the Oxford English Dictionary to be a contraction or stricture of a passage, duct or canal. The term originates from the Greek term stegnosis, meaning narrow (OED-Online 2019b). The term was prominent in the writings of ancient medical authors but later largely disappears from the medical literature until the nineteenth century when it changed subtly in meaning to become a term used primarily for the description of impediment to the transport of fluids through tubular structures, used particularly when discussing vascular narrowing (Verbiest 1976).

While Verbiest himself popularised "stenosis" for the description of canal narrowing, he later considered this a mistake, having intended the term to refer to the observed obstruction of contrast flow on myelography rather than directly to the anatomical changes. Having switched to the term narrowing for subsequent papers, he eventually admitted defeat, and switched back to using lumbar canal stenosis in 1976 (Verbiest 1976), submitting to the trend he had started.

The first attempt to provide a consensus definition of lumbar spinal stenosis was provided by Arnoldi et al. (1976). The definition included any type of narrowing of the spinal canal, nerve root canals or intervertebral foramina and allowed it to be local, segmental or generalised; caused by bone or soft tissue; and involve the canal, dural sac or both. The definition also included narrowing caused by disc herniations. Verbiest objected to this definition, believing it too broad, and thought the term should be limited to cases where the whole content of the canal was compressed between two opposite parts of the canal wall (Verbiest 1976).

The literature's use of the term stenosis remains vague, as discussed in the previous section. While a 2013 consensus definition by the North American Spinal Society (NASS) has helped a little (discussed further in § 1.5.1), it is not universally accepted and has its own problems. Part of the issue is that despite advancements in diagnostic technology, understanding of the link between the anatomical changes and the pathophysiology of the condition remain poorly defined and the debate between the differential contributions of cauda equina compression and other potential pathological mechanisms continues (Andreisek et al. 2011).

The 1970s and 1980s saw the development of computed tomography (CT) and MRI imaging which revolutionised the diagnostic approach to spinal conditions. Verbiest began work on CT imaging in LSS by the late 1970s (Verbiest 1979), however, as access to cross-sectional imaging become easier and more prevalent the focus of the literature began to shift away from the congenital and developmental causes of stenosis he had described, towards the degenerative causes of canal narrowing that it rapidly became obvious was much more prevalent (Brinjikji et al. 2015b). The diagnostic potential of one of these new imaging modalities, MRI, for the diagnosis of anatomical LSS will be explored in this thesis.

1.3 Chordata

The anatomy of the lumbar spinal canal, its normal function and variability within and between individuals.

1.3.1 Basic anatomy of the lumbar spine

The spine is a complex biomechanical system, consisting of a variable number of bony vertebrae and the articulations between them (Bogduk 2012). The spine serves to support the trunk and head, acts as a site of muscular attachment, and also protects the spinal cord and nerve roots that ultimately form the somatic and sympathetic nervous supply to the trunk and limbs. The lumbar spine normally consists of five vertebra and runs between the sacrum, which connects the spine to the pelvis, and the thoracic spine. The lumbar vertebrae are labelled L1 to L5, moving from superior to inferior levels.

The anterior aspect of each vertebra consists of a box shaped vertebral body that serves the primary weight bearing function of the spine. Load is transmitted from vertebral body to vertebral body in the longitudinal direction through the superior and inferior end plates, forming the roof and floor of each vertebral body box. The internal structure of the vertebral body resists this loading through a lattice of trabeculae: longitudinal trabeculae transmit load between the superior and inferior end-plates and horizontal trabeculae prevent any tendency of the vertical struts to bow under load (Izzo et al. 2013a). Moving inferiorly in the spine, there is a progressive increase in vertebral body size corresponding to the increasing load each inferior vertebral body must carry (Fang et al. 1994). This trend ends at the sacrum.

The end-plates of the vertebral bodies are separated from their neighbours by cartilaginous intervertebral disc (Clarençon et al. 2016)—this separation is necessary to facilitate rocking movements of the vertebral bodies relative to each other and hence allow flexion and extension movements (Bogduk 2012). Similar to the vertebral bodies, the discs increase in size moving inferiorly, reaching a maximum at the L4–5 disc (Fang et al. 1994).

Centrally the disc consists of a gelatinous nucleus pulposus, semi-fluid in consistency, that is surrounded and kept in position by the annulus fibrosus (F. Wang et al. 2016). Embryologically the nucleus pulposus is formed from the notochord (Zhao et al. 2007) and around 65% of its dry weight consists of proteoglycans. The high proteoglycan content results in retention of water in the disc that constitutes about 70% to 90% of its weight *in vivo* and gives it its fluid properties (Bogduk 2012). The annulus consists of 10 to 20 concentric lamellae and is strongly attached to the margins of the end-plates of the adjacent vertebral bodies. Each lamella is only loosely interconnected to its neighbours, allowing the disc to deform during movement of the spine (Bibby et al. 2001). Under axial loading, the nucleus pulposus exerts pressure in all directions, transferring the majority of the load between vertebrae, and is prevented from escaping the disc space by the annulus and the end-plates. Intradiscal pressures may rise on moving from a reclined to sitting position by up to 100% (Nachemson et al. 1964). The annulus stretches under such loads and hence stores energy, allowing it to act as a shock absorber. In bending movements, the annulus on the side of the disc away from the movement may also transfer tension between vertebrae (Bibby et al. 2001). The end-plates are cartilaginous and cover the superior and inferior aspects of the disc. They are composed of hyline cartilage centrally and fibrocartilage at the site of articulation with the inner aspect of the annulus fibrosus. (Bogduk 2012)

Posterior to each vertebral body extends an arch of bone—the neural arch. The base of the arch bilaterally is formed by the pedicles that extend posteriorly from the posterolateral aspects of the vertebral body. From the posterior aspect of the pedicles the lamina extend towards the midline, connecting to complete the arch. Three elongated bony processes project from the arch: from the posterior midline the spinous process, and laterally from the junction of the pedicles and laminas paired transverse processes. These processes act as levers for various muscular and ligamentous attachments (Bogduk 2012).

The lamina also forms two articular processes on each side, inferior and superior, which extend towards the articular processes of the adjacent neural arches and form faceted diarthrodial synovial joints (Clarençon et al. 2016). These joints are variously called zygapophyseal joints, apophyseal joints or facet joints (a non-specific term Professor Bogduk (2012) calls “lazy and deplorable”). The joints are surrounded by a capsule that is replaced anteriorly by the ligamentum flavum. The part of the lamina between the superior and inferior articular process is known as the pars interarticularis and the cortical bone there is generally thicker any other part of the lamina (Beresford et al. 2010), reflecting its role in transmission of forces during loading of the facet

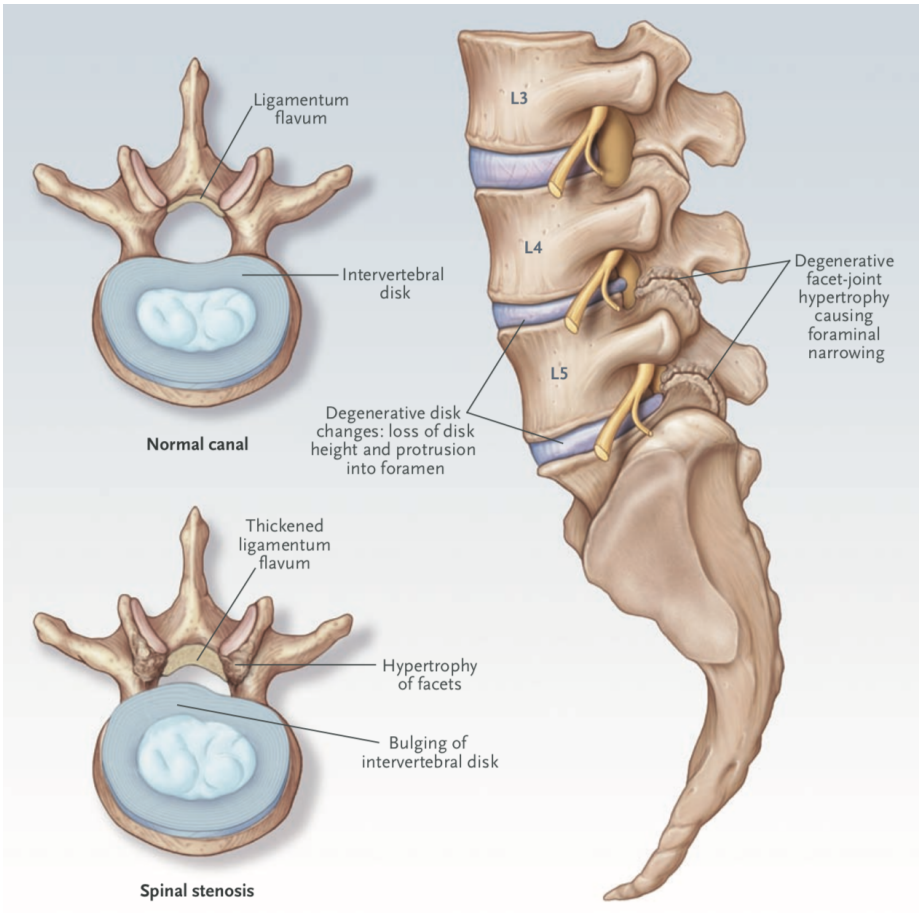


Figure 1.1: Basic anatomy of the lumbar vertebral column with demonstrated of degenerative changes that may occur with age. Reproduced with permission from Katz (2008). Clinical practice. Lumbar spinal stenosis. The New England Journal of Medicine, 358(8), 818–825. Copyright Massachusetts Medical Society.

joints.

The zygapophyseal joints support about 16% of the vertical load in normal standing posture but in extension the compressive load increases as the facet joints close (Dreyer et al. 1996). They also play major roles in stabilising the spine during movement. During sudden application of a shear force both the intervertebral disc and zygapophyseal joints resist, however, if the force is applied slowly the viscoelastic nature of the disc and horizontal orientation of the end plates means the majority of this work falls to the zygapophyseal joints (Serhan et al. 2007). The types of shear and rotational forces resisted depends on the joints' orientation and this can vary from level to level to restrict or allow relative movement of adjacent vertebrae in different ways. In the superior lumbar levels the facets are more sagittal and curved in orientation, similar to those of the thoracic spine, helping resist axial rotation. There is a gradual transition moving inferiorly towards a more coronal and flat orientation, helping resist against flexion and antero-posterior shear (Beresford et al. 2010). Similarly, there is a general increase in surface area of the articular facets moving from L1 to S1 (Serhan et al. 2007), likely reflecting the increasing load the lower vertebra carry.

Passive stabilisation of the spine is provided by a number of ligaments. The anterior and posterior longitudinal ligaments connect the anterior and posterior surfaces of the vertebral bodies as a long band, inferiorly connected to the sacrum and superiorly continuing over the thoracic vertebra (Loughenbury et al. 2006). The ligaments are connected both to the superior and inferior aspects of the vertebral body surfaces and to the anterior and posterior disc annulus. They are not attached to the concave mid vertebral body and create a small space deep to the ligament at each vertebral level. The posterior longitudinal ligament, broad over each disc, narrows to a thin strip in its unattached mid-vertebral segments.

Multiple ligaments also join the neural arches of each vertebra. The ligamentum flavum is a broad band, joining the lamina of each vertebra and forming the anterior capsule of the zygapophyseal joints. The ligament consists of around 80% elastin (giving it the yellow colour for which it is named) and 20% collagen. Its elastic nature is likely to be important in helping restore the spine to a neutral from extended position, where the neural arches are separated and the flavum stretched (Grenier et al. 1987). Between each spinous process runs fibres of the interspinous ligament, these fibres are largely parallel to the processes and hence provide little resistance to flexion. In addition, a supraspinous ligament runs over the ends of the spinous processes at superior levels, terminating in most individuals at L3. Sheets

of connective tissue also run between the transverse processes, are often called inter-transverse ligaments, but are in reality more of a fascial plane separating the anterior from posterior spinal musculature (Bogduk 2012).

The normal spine demonstrates three curvatures, a cervical and lumbar lordosis and a thoracic kyphosis, forming an overall S-shape in the sagittal plane. These curvatures play a role in increasing resistance to vertical loads by directing loading forces into the muscular contractions required to maintain the curvatures (Galbusera et al. 2016). The lumbar lordosis is in part determined by pelvic factors—the sacrum is tilted so that its upper surface that articulates with L5 slopes downwards and forwards. In order to regain an upward orientation the lumbar spine must therefore curve posteriorly. The relatively posteriorly wedged shaped nature of both the L5 vertebra and L5/S1 disc helps achieve this. The lordosis is not fixed but reduces with forward flexion and increases with extension movements (Bogduk 2012).

Having described the basic structure of the lumbar spine, we will now move on to describe in detail the lumbar spinal canal and its contents, whose narrowing with age largely forms the subject of this thesis.

1.3.2 The lumbar spinal canal and its contents

The boundaries of the canal

The space encircled by the neural arch of each vertebra is referred to as the vertebral foramina. The lumbar spinal canal is the channel formed by vertebral foramina at each lumbar level and the ligamentous structures joining them. Its function is to conduct and protect the nerve roots of the cauda equina as they run from the termination of the spinal cord to their exit from the canal, where they ultimately form the nerves that innervate the lower back, pelvis and lower limbs (Bogduk 2012).

As is implied by the above, the canal has alternating bony and ligamentous portions (Wilmink 2010b). The bony segments are formed by the neural arch, with the vertebral body anteriorly, the pedicles laterally and the lamina posteriorly. The ligamentous portion is formed by the posterior disc annulus anteriorly (with overlying posterior longitudinal ligament), the ligamentum flavum posteriorly and facet joint capsules posterolaterally. Inferior to each pedicle the canal is open (the neural exit foramen) and this opening allows both the nerve roots to exit and for arterial, venous and lymphatic supply of the canals contents.

The normal shape of the canal changes from oval at L1 to a triangular shape at L5 and S1 (Aly et al. 2013). A trefoil shaped canal, while often associated with enlargement of the facet joints, is seen in around 30% of asymptomatic individuals at L5 and is likely to represent a normal developmental variant (Gouzien et al. 1990). Discussion of normal canal size is made, alongside descriptions of the various imaging measurements of the canal, in §1.3.3.

The cauda equina

The spinal cord ends inferiorly at the conus medullaris, which normally lies at the L1–2 vertebral level. From the conus arise multiple ventral (sensory) and dorsal (motor) rootlets that coalesce to form first nerve roots and finally the mixed spinal nerves that provide the segmental supply of the pelvis and lower limbs. The spinal nerves are labelled according to the neural exit foramen through which they leave the canal, generally one spinal nerve per foramen, and hence are labelled L1 through L5 for the lumbar spinal nerves and S1 through S5 for the sacral spinal nerves. The nerve roots within the canal are are collectively referred to as the cauda equina (*horse's tail*) (Wilmink 2010b).

The spinal cord and cauda equina is surrounded by an extension of all three cranial meningeal layers. The structure of the spinal meninges largely conforms to that of the cranial meninges, with the exception that the dura mater is not directly adherent to the bony boundaries of the canal and instead forms the dural or thecal sac (Vandenabeele et al. 1996). The sac contains the cerebrospinal fluid (CSF) in which the cord and cauda equina bathe and normally ends at the upper part of the S2 vertebral body (Bogduk 2012). It is surrounded by an epidural space, described in the section below.

Within the normal dural sac, the cauda equina roots settle dependently, layering posteriorly in a supine individual. Moving inferiorly, nerve roots about to leave the canal first move from the central canal to the lateral margins of the dural sac before leaving the dural sac proper within a nerve root sleeve. The sleeve constitutes a roughly cylindrical extension of dura mater and CSF and passes through the lateral aspect of the canal en-route to the neural exit foramen (Wilmink 2010b). The dorsal root ganglion, a swelling of the dorsal nerve root which contains sensory neuron cell bodies, lies at the termination of the nerve root sleeve within the neural exit foramen, usually directly inferior to the pedicle.

In a cadaveric study of 33 individuals, Suh et al. (2005) demonstrated that the nerve root sleeves at all lumbar levels leaves the dural sac superior to the disc level at which the nerve root exits and hence travel obliquely towards the neural exit foramen. The sleeves originated just below the pedicle in the upper lumbar spine, giving a shorter and more more transverse course relative to inferior lumbar levels where the sleeve originated closer to the disc level above and hence had a longer and more longitudinal course. The L5 nerve root sleeves were sometimes seen to originate above the level of the L4/5 disc.

The dural sac is partially tethered to the surrounding canal by concentrations of the surrounding epidural fascia (meningovertebral ligaments), ventrally this takes the form of a network of filaments and median septa to the posterior longitudinal ligament. In addition, each nerve root sleeve is both tethered anteriorly to the posterior longitudinal ligament and laterally to the pedicle (Barbaix et al. 1996). The nerve

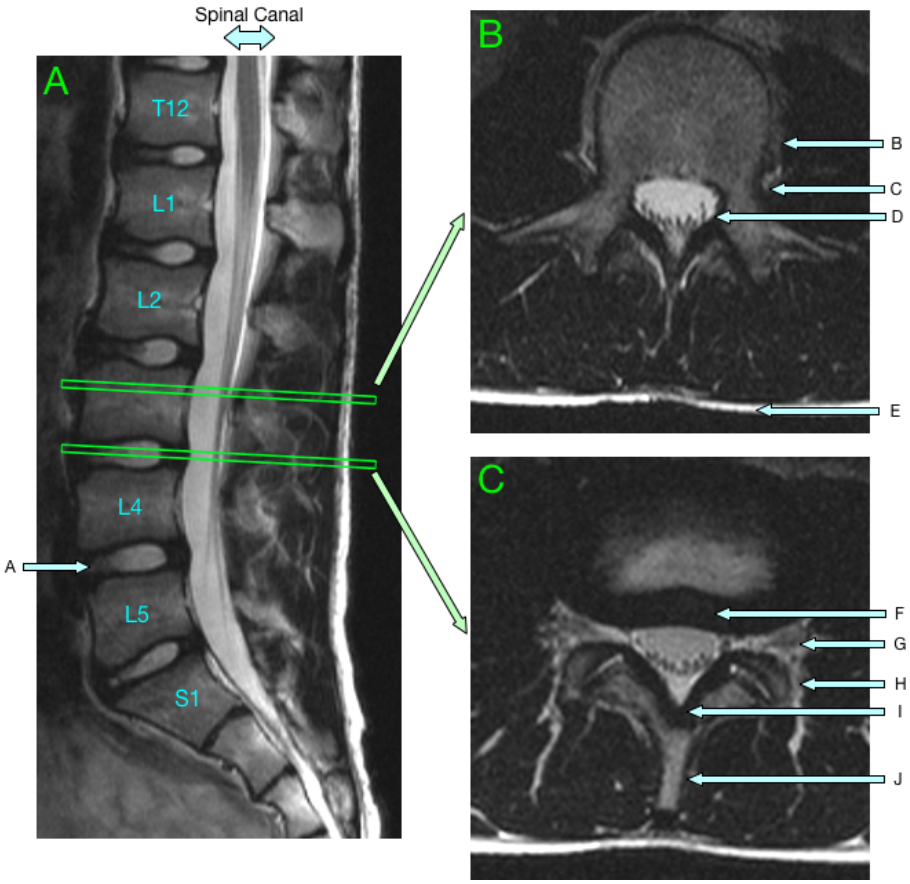


Figure 1.2: Anatomy of the spinal canal demonstrated on T2-weighted MRI. Panel A: Mid-sagittal image with labelled vertebral bodies, excluding L3. Panel B: Axial slice at the pedicular level. Panel C: Axial slice at the disc level. Arrowed structures: A – intervertebral disc, B – vertebral body, C – pedicle, D – nerve roots within the spinal canal, E – skin of the back, F – posterior annulus of the intervertebral disc, G – nerve root within the neural exit foramen, H – facet joint, I – ligamentum flavum, J – spinous process.

root sleeve is also bound by four foraminal ligaments extending in the form of a cross to the bony boundaries of the neural exit foramen (Grimes et al. 2000). Despite these attachments, the dural sac and nerve root sleeves can show considerable movement during flexion and extension of the spine (Wilmink et al. 1988).

Once the spinal nerve roots have formed and left the canal they divide forming a dorsal ramus, which innervates the local spinal structures and skin of the back, and a ventral ramus that contributes to the lumbosacral plexus (Bose et al. 1984).

The lateral recess and neural exit foramen

The lateral recess is the part of the canal containing the exiting nerve root and its dural sleeve as it moves towards the neural exit foramen. It has both a superior ligamentous portion and an inferior bony portion. The ligamentous lateral recess is formed by the posterolateral disc annulus anteriorly and posteriorly by the zygapophyseal joint capsule/ligamentum flavum. The inferior bony lateral recess is formed by the dorsal surface of the vertebral body anteriorly, the superior articular process of the zygapophyseal joint posteriorly and the pedicle laterally. The medial aspect of both bony and ligamentous portions is open and defined by a plane running anteriorly from the medial edge of the superior articular facet (Wilmink 2010b). While the definition of the lateral recess is somewhat arbitrary, its importance comes from the fact degenerative hypertrophy of the ligamentum flavum and zygapophyseal joint can cause it to be significantly narrowed, entrapping the nerve root transiting to the foramina and causing radicular pain, and this can occur in isolation from any degenerative change affecting the rest of the canal (Bartynski et al. 2003). Wilmink (2010b) suggests the true lateral recess only exists in the lower lumbar spine where the canal becomes triangular or trefoil in morphology. In the upper lumbar spine, little of the canal lies below the articular processes.

The neural exit foramen forms the lateral aspect of the spinal canal inferior to the pedicles at each level bilaterally. It lies at the inferior aspect of bony lateral recess and is bordered both superiorly and inferiorly by the pedicles of the adjacent vertebra. Anteriorly the superior aspect is formed by the posterior vertebral body cortex and the inferior aspect is formed by the posterior disc annulus. The posterior border of the canal is formed by the articular processes and the overlying ligaments (Bose et al. 1984). The foramina are normally round or oval in shape, with cross sectional area increasing between L1/2 to L4/5

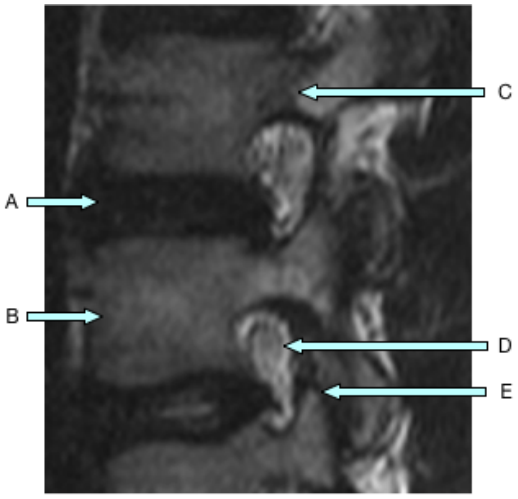


Figure 1.3: Sagittal anatomy of the neural exit foramen on T2-weighted MRI sequences. Arrowed structures: A – intervertebral disc, B – vertebral body, C – pedicle, D – nerve root within the neural exit foramen, E – superior articular facet.

with the L5/S1 foramina being much smaller than the rest (but also containing the largest nerve root) (Bogduk 2012). The spinal nerve passes through the superior aspect of the foramen with ligamentous attachments as already described. Several transforaminal ligaments have also been described, the most common being a band running from posterolateral vertebral body to the base of the transverse process of the same vertebra (Bogduk 2012).

Some authors consider the bony lateral recess and neural exit foramen together to constitute a nerve root canal, which, like the nerve root sheathes, become longer and more longitudinally orientated at inferior levels (Bose et al. 1984).

The epidural space, vascular and nervous supply of the canal

The epidural space surrounds the dural sac and contains adipose tissue, and a plexus of veins, arteries and nerves, all contained within loose areolar connective tissue (Bogduk 2012). The epidural space anterior to the dural sac is divided behind each vertebral body by the posterior longitudinal ligament. Thin membranes extend bilaterally from the posterior longitudinal ligament to insert into the lateral walls of the spinal canal, forming an anterior epidural space behind each vertebral body, separate to the rest of the canal (Schellinger et al. 1990). This space is divided in two by a membrane in the midline connecting the posterior longitudinal ligament to the vertebral body (Loughenbury et al. 2006).

Epidural fat partially or completely surrounds the dural sac. It often present anteriorly, smoothing the irregular contour of the posterior vertebral bodies and present posteriorly as a retrodural fat pad, between the dural sac and ligamentum flavum. The size of the retrodural fat pad tends to increase at more inferior lumbar levels (Wilmink 2010b).

The arterial supply of the canal contents is derived from the paired lumbar arteries that arise from the aorta ventral to each lumbar vertebra and wrap around the vertebral body bilaterally. At the level of the neural exit foramen the arteries divide to form a posterior branch that anastomoses around the posterior aspect of the neural arch and several branches that enter the intervertebral foramina. An anterior spinal canal branch runs along the anterior aspect of the canal, anastomosing between levels, a posterior spinal canal branch runs along the inner aspect of the lamina and ligamentum flavum and a radicular branch forms a plexus around the dorsal root ganglion before entering and running superiorly in the nerve root as the distal radicular artery (Bogduk 2012).

The venous drainage of the canal contents is primarily via the internal vertebral venous plexus. This anterior plexus has a ladder like appearance, connecting across the midline in the anterior epidural space behind each vertebral body, and moving laterally in both directions to pass the disc over its posterolateral aspects. No veins pass directly over the dorsal part of the disc annulus (Manaka et al. 2003). A less significant posterior intervertebral plexus also runs along the lamina and flavum. The internal spinous plexus connects to the ascending lumbar veins (part of the external vertebral plexus) via veins passing through the intervertebral foramina. These veins form part of an external vertebral plexus which drains to the common iliac veins inferiorly and the azygos system superiorly (Wilmink 2010b).

The nerve roots of the cauda equina receive a dual blood supply and drainage. Proximally they are supplied by dorsal and ventral proximal radicular arteries (arising indirectly from the posterior and anterior spinal arteries) which enter each root, join to form the proximal radicular artery, and follow the root along its length before anastomosing with the distal radicular artery derived from the foraminal branches of the paired lumbar arteries. Similarly drainage of the nerve roots occurs by proximal and distal radicular veins, draining towards both the conus and veins within the neural exit foramina (Bogduk 2012).

The primary innervation of canal structures is

via sinuvertebral nerves that branch from the ventral rami and re-enter the canal through the intervertebral foramina (Bogduk 2012). These form an ascending branch, that innervates the posterior longitudinal ligament (PLL) and disc at the level above, and a descending branch, that supplies the disc at the level of the foramen. A plexus of nerves surrounds the anterior aspect of the dural sac and also provides autonomic innervation of the various canal vessels. The disc themselves have been shown to contain a variety of free and complex nerve endings in the outer third of the annulus (S. Roberts et al. 1995), providing both proprioceptive and nociceptive supply. Similarly, the facet joints receive nerve fibres from branches of the dorsal rami that run across the superficial aspect of the lamina. The facet joint capsule is richly innervated with nociceptive, proprioceptive and autonomic fibres (Kalichman et al. 2007).

1.3.3 The size of the normal adult canal

The following section presents the available information regarding the size of the normal adult lumbar spinal canal. Measurements relevant to cross-sectional imaging are discussed (figures 1.4 and 1.5) with tabulation of normal values where available. Evidence from both CT and MRI studies is included, as while direct comparisons of quantitative measures of canal size between CT and MRI tend to show statistically significant disagreements, the absolute difference in measurements between imaging modalities tends to be small (Eun et al. 2012), and the available evidence for MRI alone is limited. Only studies of populations confirmed to be asymptomatic are included.

The presented normal values differ from study to study, likely in part due to differences in their study populations, however, it is also important to note that there is considerable variability in terms of the exact methodology used for producing each measurement. Differences in measurement technique extend from the orientation of the images used for the measurement through to the name used to describe the measurement. While effort has been taken to tabulate some of the likely causes of variability in the presented normal values, this is unlikely to be complete and hence caution is required in the tables interpretation.

The tabulated evidence on normal values is compiled from the following sources:

- The list of evidence on normal canal size provided by Wilmink (2010c).
- The list of papers relating to developmental stenosis provided by J. P.-Y. Cheung et al. (2014b).
- The systematic review of quantitative measurements used in LSS by Steurer et al. (2011).
- Additional citations based upon the bibliographies of the included papers and any relevant evidence found while writing this thesis.

The names used for the various measurements in the following descriptions are adopted as conventions for the rest of this thesis.

Measurements of the bony canal

Measurements of the size of the vertebral foramina, which define the bony borders of the canal, were particularly popular in the early literature relating to LSS. This is most likely secondary to the limited ability of the techniques of the time to demonstrate soft tissue structures. In contrast, the bony aspects of the canal could be easily measured both through cadaveric studies (e.g. Amonoo-Kuofi 1985) and on plain radiographs (e.g. Hinck et al. 1965a). Despite the availability of newer imaging modalities that can demonstrate soft tissue structures, measurements of the bony canal continue to be commonly used in the literature. They are useful for studies based on CT imaging, a technique that has limited soft tissue resolution (Kreiner et al. 2013). Studies investigating developmental LSS also tend to use bony measurements of the canal under the assumption several of these measurements change little with age and are not affected by degenerative change, effectively making them signposts for the original size of the canal at the completion of growth (Schizas et al. 2014). This strength in the assessment of developmental stenosis is also a significant weakness for assessment of degenerative pathology that tends to affect the disc and ligamentous structures. In the presence of degeneration these measurements may dramatically misrepresent the actual amount of remaining space available for the canal contents (Schonstrom et al. 1985).

The *antero-to-posterior canal diameter (APD)* is measured between the anterior aspect of the base of the spinous process, in the midline where the lamina join, and the posterior aspect of the vertebral body. It was originally described as a measurement made on lateral radiographs of the lumbar spine (Hinck et al. 1965b) but has been adapted for use in cross-sectional imaging, using either midline sagittal images (figure 1.4 measurement 10) or a transverse section taken at the level of the pedicles (figure 1.4 measurement 6). Studies using sagittal images tend to either record the shortest APD per spinal level or alternatively measure consistently at the mid-vertebral body.

The measurement is of particular historic importance, being the first measurement used to define developmental LSS on radiographs (Verbiest 1976). Consistent with its use as a marker for the developmental growth of the canal, R. W. Porter et al. (1980) showed that mean canals only differed by 1 mm between a group of 15 to 21 year old and a second group of 50 to 65 year olds. Similarly, H. M. Lee et al. (1995) found no significant relationship between the APD and age in a cadaveric study of adults.

The collected evidence on the normal values for the APD is presented in tables A.1 and A.2. A variant of the APD, the oblique APD, has been used in ultrasound (US) imaging of the spinal canal. US is unable to measure the APD directly due to poor

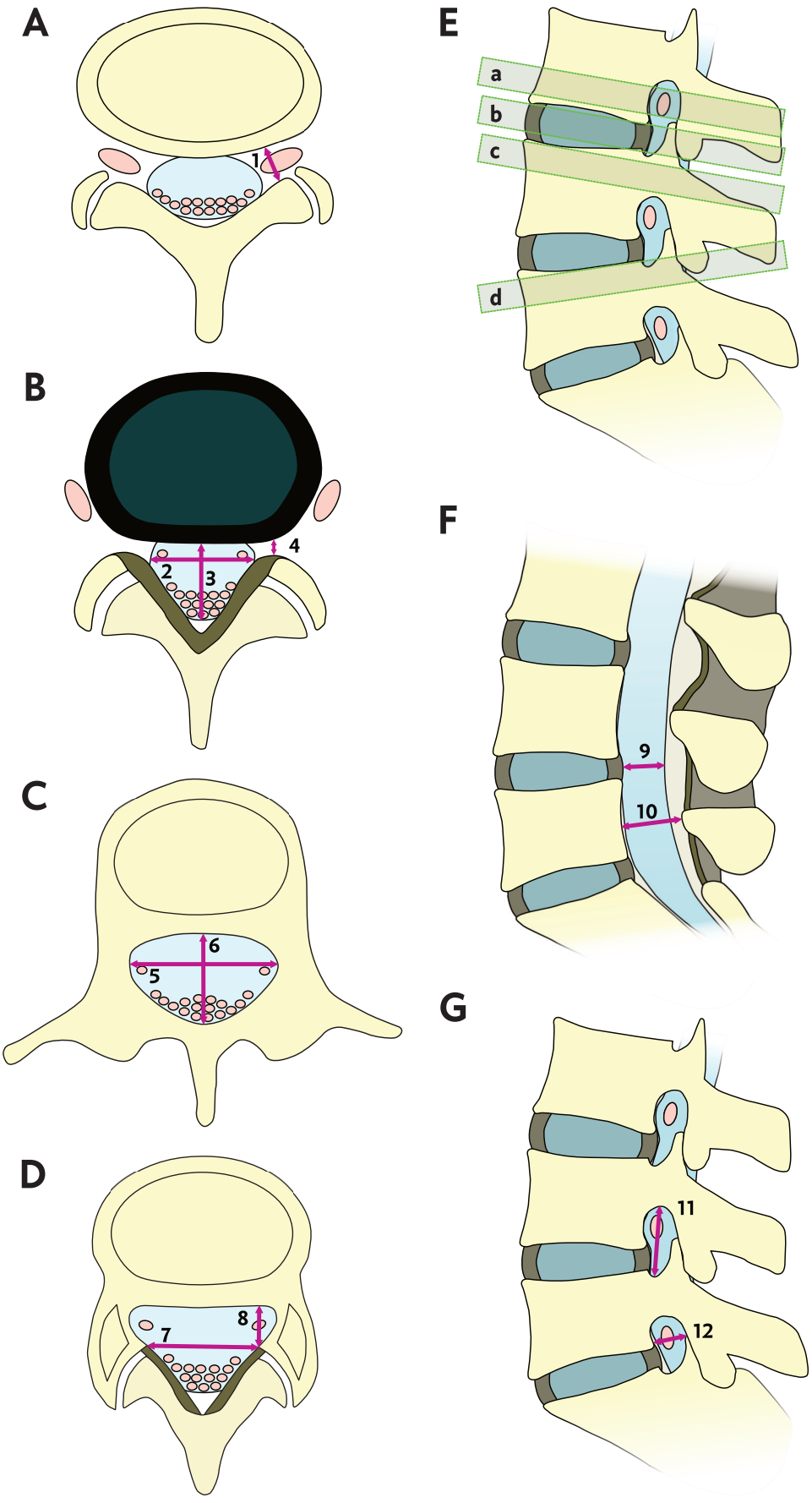


Figure 1.4: Measurements of the lumbar spinal canal. Panels A-D: cross-sections at planes indicated on panel E. Panel F: mid-sagittal section. Measurements: 1 – neural exit foramen diameter (NFD), 2 – dural-sac transverse diameter (DS-TD), 3 – dural-sac antero-to-posterior diameter (DS-APD), 4 – sub-articular sagittal diameter of the canal (SSDC), 5 – inter-pedicular canal diameter (IPD), 6 – antero-to-posterior canal diameter (APD), 7 – inter-facet canal diameter (IFD), 8 – lateral recess depth (LRD), 9 – dural-sac antero-to-posterior diameter (DS-APD), 10 – antero-to-posterior canal diameter (APD), 11 – neural exit foramen height (NFH), 12 – neural exit foramen diameter (NFD).

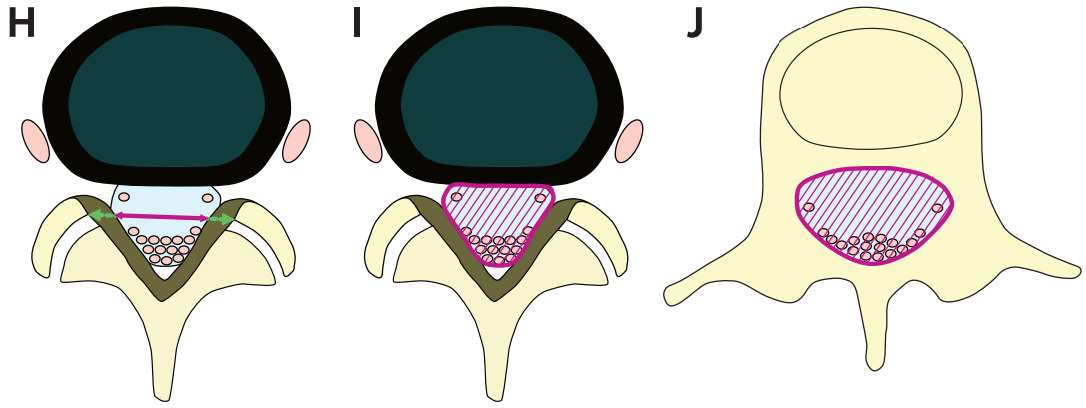


Figure 1.5: Additional measurements of the lumbar spinal canal. Panel H: cross-section at the disc level, demonstrating the interligamentous canal diameter (ILD) (purple) and inter-facet canal diameter (IFD) (green hashed + purple). Panel I: cross-section at the disc level showing the dural-sac cross-sectional area (DS-CSA) (purple hashed area). Panel J: cross-section at the pedicular level showing the vertebral foramen cross-sectional area (VF-CSA) (purple hashed area).

transmission of ultrasound waves through the spinous processes (R. W. Porter et al. 1978).

The *inter-pedicular canal diameter (IPD)* was originally developed as a measurement on anterior-to-posterior (AP) radiographs of the lumbar spine (Hinck et al. 1966) and is measured as the minimum distance between the inner aspects of the pedicles at each vertebral level. The measurement was adapted for use on cross-sectional imaging where it is generally measured on axial slices at the pedicular level, taking the maximum transverse diameter between the pedicles (figure 1.4 measurement 5). Similarly to the APD, it is of importance in characterising developmental stenosis as it is relatively fixed after completion of growth (H. M. Lee et al. 1995), but has generally been found to be less commonly narrowed in developmental stenosis than the APD (Eisenstein 1977). The evidence relating to normal values of the IPD is presented in tables A.3 and A.4.

The *inter-facet canal diameter (IFD)* is measured between the most medial aspects of the two facet joints on transverse images, generally at the level of the disc (e.g. Gouzien et al. 1990, figure 1.5 H), but also can be measured at the infra-pedicular level (e.g. Kornberg et al. 1985, figure 1.4 measurement 7). Unlike the two previous measures of the bony canal, this measurement seems much more likely to change with facet joint degeneration and associated hypertrophy, particularly as degenerative osteophytes may develop at the medial joint line (Kalichman et al. 2007). The evidence relating to normal values of the IFD is presented in table A.5.

The *vertebral foramen cross-sectional area (VF-CSA)* is measured on transverse images of the lumbar spine taken at the pedicular level (figure 1.5 J). Unlike the previously described measurements its use generally requires computer software, which is used to draw a region of interest with the same internal borders as the vertebral foramen: the posterior vertebral body cortex and inner aspects of the pedicles and lamina. The measurements value is likely to be closely related to the APD and IPD at the same level, which make up the measured areas AP and transverse diameters respectively. The available evidence for normal values of the VF-CSA is presented in table A.6.

Measurements of the ligamentous canal

Only a single measurement of canal size incorporating soft tissue structures was identified—the *interligamentous canal diameter (ILD)*. This measurement is made between the inner aspects of the ligamentum flavum on each side of the canal, directly medial to the facet joint line (figure 1.5 H). The measurement is made on transverse images at the disc level and is very similar to the IFD, being shorter than the corresponding IFD by the combined thickness of the ligamentum flavum overlying the facet joint lines. Also similar to the IFD, the ILD is likely to become smaller with accumulating degenerative change as the facet joints and ligamentum flava undergo hypertrophy. Given this, it perhaps not surprising that in a study of 50 individuals referred for suspected lumbar nerve root compression by degenerative pathology, the measured ILD was significantly smaller than in a control group, while the APD and IPD showed no significant difference between groups (Wilmink et al. 1988). The available evidence relating to normal values of the ILD is presented in table A.7.

Measurements of the dural-sac

Given the potential role of nerve root compression in the pathophysiology of symptomatic LSS, it is perhaps not surprising that measurements of the dural sac are common in the literature. As the canal gets smaller, the dural sac is expected to become smaller as it undergoes compression, and measurements of the dural sac provide a more direct assessment of the space available for the cauda equina. The dural sac can be visualised *in vivo* on various imaging modalities, including: myelography (Amundsen et al. 1995), plain CT (though with reduced confidence due to relatively poor soft tissue contrast (Schonstrom et al.

1985)), CT myelography and MRI (Schnebel et al. 1989).

The *dural-sac antero-to-posterior diameter (DS-APD)* and *DS-TD* are measurements of the dural-sac at its greatest diameters in the AP and transverse directions. The former may be measured on both mid-sagittal (e.g. Chatha et al. 2011, figure 1.4 measurement 9) and transverse images (figure 1.4 measurement 3), while the latter is measured on transverse images alone (figure 1.4 measurement 2). The available evidence for normal values of the DS-APD and DS-TD are presented in tables A.8 and A.9 respectively.

The *dural-sac cross-sectional area (DS-CSA)* is the most commonly used quantitative measurement in studies of LSS (Steurer et al. 2011), its popularity likely reflecting its ability to reflect narrowing imposed from all sides, unlike the DS-APD or DS-TD, which could potentially could remain normal if the dural-sac was compressed in the opposite plane to the measurement. Similar to the VF-CSA it is generally measured using computer software by drawing a region of interest contained by the dural-sac boundaries on transverse images (figure 1.5 I), though an alternative is to measure the DS-APD and DS-TD and calculate the DS-CSA, adjusting for its shape at the site of measurement (Hamanishi et al. 1994), both methods show comparable results.

One limitation of the DS-CSA is its dependence on slice orientation, with a 10% change in slice orientation resulting in a change in DS-CSA by a mean of 5% (Henderson et al. 2012). This is more likely to be a problem when axial images are taken as a single block rather than as separate blocks angled to the disc. Some papers have attempted to adjust for this, by measuring the slice angle relative to the dural-sac and adjusting the measured DS-CSA accordingly (Hamanishi et al. 1994). This limitations is not likely limited to the DS-CSA but likely affects all quantitative measurements of the canal in same orientation as which the slice angle is being changed.

Measurements of the lateral recess

In a 2011 systematic review of quantitative measurements of the lumbar canal, Steurer et al. (2011) identified three quantitative measurements of the size of the lateral recess: the *lateral recess height (LRH)*, *lateral recess depth (LRD)* and *lateral recess angle (LRA)*.

The LRH was first described by Ciric et al. (1980) as the distance between the most anterior aspect of the superior articular facet and the posterior border of the spinal canal formed by the posterior vertebral body cortex (figure 1.4 measurement 8). This measurement was taken at the level of the pedicle, across the bony portion of the lateral recess, and was described on both radiographs and CT imaging.

The LRD was described one year later by Mikhael et al. (1981) where it was described as a measurement between the antero-medial aspect of the superior articular facet anterior to the posterior vertebral body cortex. It was again measured on CT at the level of the pedicle.

Despite the description of the LRH and LRD as separate measurements within Steurer's systematic review, the exact difference in measurement technique between the two measurements is unclear both from the systematic review text or figures. Inspection of the papers cited as defining the measurements also fails to distinguish the measurements. At the level of the pedicle, the anteromedial portion of the superior articular facet is also the most anterior aspect. In fact, both Ciric and Mikhael were authors on both the papers defining these measurements, do not attempt to describe any difference between the measurements, despite the differing measurement name between papers, and hence seem to have considered LRD and LRH to be synonyms. Other papers describing the bony LRD or LRH have also not made a clear distinction between the measurements (e.g. Strojnik 2001; Dincer et al. 1991; Santiago et al. 2001; Aly et al. 2013). For the remainder of this thesis the measurement will be referred to as the LRD, as this is the most commonly used term within the literature. The available evidence on normal values of the LRD is presented in table A.11.

An MRI variant of the LRD was described by Chung et al. (2000) measuring the ligamentous rather than bony lateral recess and described as the SSDC. The measurement is made at the level of the disc, from the most anterior aspect of the ligamentous/capsular coverings of the superior articular facet anterior to the posterior disc annulus in the sagittal plane (figure 1.4 measurement 4). Due to changing morphology of the zygapophyseal joints with spinal level and associated degeneration, the plane of this measurement varies from lying immediately anterior to the most medial aspect of the superior articular facet (and hence across the most medial aspect of the lateral recess) to a position much more laterally within the canal, effectively crossing the inferior aspect of the neural exit foramen. The available evidence on normal values of the SSDC is presented in table A.11.

The *lateral recess angle (LRA)* is defined as the angle between lines running along the floor and roof of the lateral recess (Dincer et al. 1991). No data on normal values was available for the LRA.

Measurements of the neural exit foramen

The *neural exit foramen diameter (NFD)* is a measurement taken in the AP direction between the most posterior and most anterior aspect of the neural exit foramen. It can be measured on sagittal images (e.g. Cramer et al. 2003, figure 1.4 measurement 12) or on axial images (e.g. Santiago et al. 2001, figure 1.4 measurement 1). The available evidence for the normal value of the NFD is presented in table A.12.

The *neural exit foramen height (NFH)* is measured from the most posterior to most superior aspects of the neural exit foramen and is measured in the sagittal plane (e.g. Cramer et al. 2003, figure 1.4 measure-

ment 11). No data fitting the criteria set out in the beginning of this section is available for its normal values.

1.3.4 Variation in canal size with body size, gender and ethnicity

The size of the spinal canal appears to be related to an individuals height — taller individuals have larger canals, predominantly in the coronal plane. Gouzien et al. (1990), in a study of 34 young adult males, found both the VF-CSA and IPD showed moderate correlation with participant height but the APD showed no such correlation. Similarly increasing height is associated with a larger neural exit foramen, in terms of both its NFD and NFH (Cramer et al. 2003). Not all studies have replicated these findings, for instance Santiago et al. (2001) found no relationship between numerous measurements of canal size and the participants weight or height. This latter study predominantly looked at individuals with low back pain and hence the presence of degenerative pathology may have obscured any underlying correlation.

The available evidence for whether the normal canal varies between men and women is unclear. In children, both radiographic and MRI studies show no significant differences in APD or IPD between the genders within age groups (Hinck et al. 1966; Hinck et al. 1965b; Knirsch et al. 2005). R. W. Porter et al. (1978) using ultrasound found smaller canals in young male minors aged 15 to 18 compared to nurses of the same age, though the differing occupations between groups may indicate other confounding variables. In adults the evidence is more mixed, with some studies finding a larger canal in men (Schizas et al. 2014; Fang et al. 1994), some finding no difference in size (Winston et al. 1984; Greenough et al. 1992) and some finding larger canals in women (R. W. Porter et al. 1980). While the results in studies of children may be affected by differential rates of growth between the genders, this effect is likely to be small given the early age at which the majority of canal growth is complete (see §1.4.1). Studies in older adults are likely to be confounded by differing rates of accumulating degenerative change—for instance degenerative spondylolisthesis is more common in women than men (Cowley 2016). Overall, it seems likely that if significant differences in the canal size between genders are truly present they are likely to be small.

The evidence for variation in canal size between individuals of different ethnicities is much clearer. Amonoo-Kuofi (1985) found significantly larger canals in Nigerian adults compared to individuals from South African or America, as reported by Eisenstein (1977) and Hinck et al. (1966) respectively. Similar differences have been found by other authors, including smaller canals in Indian skeletons compared to Italian skeletons (Postacchini et al. 1983), and larger canals on CT scans of Caucasians compared to Chinese individuals (Fang et al. 1994). It is important to note that while the presented differences are statistically significant, the absolute difference in means between the various studied populations remains small, with substantial overlap in the total distribution of measurements.

1.3.5 Variation in canal size during movement

While the bony portion of the canal remains fixed in the mid to long term, the ligamentous portion of the canal (the motion segment) can vary with movement of the lumbar spine. In flexion and extension movements, the axis of rotation between vertebrae lies in the posterior aspect of the intervertebral disc (Wilmink 2010b). With increasing extension the dorsal aspects of the end-plates come together and the annulus bulges backwards into the canal. This bulging does usually not exceed 1 to 2 mm (Penning et al. 1981). The lamina and spinous processes also come together and this movement is relatively greater than that of the end-plates due to their greater distance from the centre of rotation (Wilmink 2010b). This movement causes the facet joints capsules and ligamentum flavum to bulge into the posterior aspect of the canal, pushing the retrodural fat pad anteriorly and reducing the area available for the dural sac. One study, consistent with the elastic nature of the ligamentum flavum, found it varied in thickness with different degrees of stretch (Schmid et al. 1999). At L4/5 in the neutral position the mean ligamentum flavum thickness was 2.3 mm, decreasing to 1.8 mm in flexion and increasing to 4.3 mm in extension. This elastic nature may be important in preventing extension induced buckling of the ligament into the canal.

The overall effect of the changes above is to reduce the ligamentous spinal canal diameter during extension. Penning et al. (1981) performed a study of 40 lateral myelograms, comparing images taken in both sitting flexed and extended positions. In extension, the dural sac anterior border became indented by the bulging discs and the posterior border was displaced by the anterior movement of the epidural fat pad. A rise in CSF pressure appeared to be avoided by bulging of the dural-sac into the anterior epidural space, presumably by displacement of blood from the valveless anterior venous plexus. In addition they described longitudinal movements of the dural-sac varying between 2 and 25 mm, at odds with the ligamentous attachments between dural-sac and the surrounding canal described in the previous section.

Schmid et al. (1999) assessed changes in the DS-CSA between flexion and extension using upright MRI studies in asymptomatic volunteers with an average age of 28 years. Consistent with the findings by Penning et al. they found an average decrease in DS-CSA of 16.4% between sitting flexion and extended positions at the L4/5 disc. This was insufficient to threaten compression of the nerve roots — even at maximal extension none of the patients had a DS-CSA of less than 150 mm^2. As expected, at the level of the

pedicles the canal dural sac was invariant with position, as was the height of the bony lateral recesses. No significant change in DS-CSA was seen moving from a sitting to standing position.

The effect of gravitation loading on the dimensions of the canal is less clear cut. B. Danielson et al. (2001) performed MRIs in 43 asymptomatic individuals in the "psoas-relaxed" position — this position has the patient in the supine position with the knees bent, and is the standard position for supine MRI, designed to maximise comfort and hence reduce movement artefact. After the standard MRI the knee wedge was removed and an axial loading device applied — this can simulate gravitation loading up to around 50% of the patients weight while remaining in the supine position to allow imaging in a standard MRI scanner. An average decrease in DS-CSA of 56% was seen on axial loading, falling to below 100 mm^2 in 37% of participants. The greater degree of dynamic narrowing compared to the findings of Schmid et al may be explained by the older participant group studied, who would be expected to have a higher proportion of degenerative disc disease. In contrast Hirasawa et al. (2007) compared standing to supine MRI in 29 asymptomatic volunteers and found no significant change in the DS-CSA at L3/4 and L4/5 on standing, but an increase in DS-CSA of 23.8% at L5/S1. They reproduced the flexion-extension related dural sac changes found by Schmid et al. The differences in findings between Danielson et al and Hirasawa et al may reflect both population factors (Danielson studied an older more degeneration population) and the supine position in the Danielson study which eliminates the rise in hydrostatic pressure exerted by the CSF in a standing position. The rise in hydrostatic pressure would tend to counter any dural sac narrowing.

Posture based changes do not just affect the central canal. The ligamentous lateral recess and neural exit foramen both narrow with extension. The lateral recess is narrowed by disc bulge anteriorly and ligamentum flavum thickening posteriorly (Penning et al. 1981). In asymptomatic volunteers movement from an upright neutral to extended position causes a reduction in 23.2% in average foraminal diameter (Schmid et al. 1999) and a 15% reduction in cross-sectional area (Inufusa et al. 1996), caused by the pedicles coming together and the disc bulging into the lower aspect of the foramen. This was not, however, sufficient to cause nerve compression in asymptomatic volunteers (Schmid et al. 1999).

1.4 A dynamic structure…

The changes in the spine that occur with age; normal, pathological and somewhere in-between

1.4.1 Growth of the lumbar canal during childhood

The bony structures of the spinal canal begin to undergo ossification around week 10 post-gestation (Bogduk 2012). Primary ossification centres form in the central vertebral body and each half of the neural arch, the latter before the former. This process begins in the thoracic levels before spreading into the cervical and lumbar spine. Ursu et al. (1996) in a study of 67 human fetal specimens found that both lumbar and sacral canals grew at the same rate up to week 14 post gestation, followed by a period of differential growth that establishes the adult ratio of vertebral body sizes. Much of the spines growth is completed *in utero*, with 70% of adult spinal canal size in the sagittal plane being reached by birth.

At the time of birth the bulk of the central vertebral body is ossified, with the upper and lower surfaces still covered by cartilage. Horizontal growth of the vertebral bodies is achieved by periosteal ossification, while longitudinal growth occurs through proliferation and subsequent ossification of the cartilage plates making up the superior and inferior aspects of the vertebral bodies (Bogduk 2012). Between 6 and 9 years of age the peripheral aspects of these plates begin to ossify, forming the ring apophysis. These fuse with the vertebra between 14 and 21 years of age, signalling the end to longitudinal growth.

The structures of the neural arch undergo a similar sequence. At the time of birth the lamina, articular and transverse processes are ossified apart from their distal tips. The ossified parts of the lamina fuse within the first year of life and the spinous process also becomes ossified during this time. The pedicles are the last part of the arch to ossify, usually fusing with the vertebral bodies and lamina by 6 years of age. At puberty, secondary ossification centres begin to appear in the tips of the spinous and transverse processes. These secondary centres fuse completely by about 25 years of age (Bogduk 2012).

Hinck et al. (1966), studying radiographs of children taken in a secondary care setting, found small progressive increase in both APD and IPD with age. The most marked growth was seen between age 3 and 10, with smaller relative increases up to the age of 18. The APD had a mean of 19 mm in the age 3 to 5 years group, increasing to a mean of 22 mm in the 17 to 18 year old group (Hinck et al. 1965b). Such radiographic measurements, while of potential clinical use, pose difficulties for understanding canal growth due to changing magnification of the structures with increasing body size and associated movement of the canal relative to the radiographic plate.

Archaeological studies have generally shown the bony spinal canal rapidly reaches its adult size in early childhood, consistent with the ages at which the ossified parts of the neural arch fuse. R. W. Porter et

al. (1987) found 87% of the adult IPD was reached by 4 years of age, with the APD reaching adult levels before this. Papp et al. (1994), expanding on these findings, found the upper lumbar canal cross-sectional area reached adult sizes by 1 year of age, whereas at L4 and L5 development was much slower, reaching maturity by 4 and 6 years of age respectively. Consistent with the findings by Porter, the growth in canal cross-sectional area was associated with increases in both APD and IPD, however, after conclusion of the period in which the canal cross-sectional area increases, the APD becomes fixed while the IPD continues to show slow growth. These processes again proceed faster in the upper lumbar spine. During this later period, increases in the canal cross-sectional area are avoided, despite the increasing IPD, by a progressive change in the shape of the canal. More recent evidence partially challenges the above, with MRI based studies suggesting the APD may continue to increase up to the age of 14 years, albeit at a slow rate (Knirsch et al. 2005).

The triangular or trefoil configuration of the lower lumbar canal appears to develop during adolescence. The canal is almost uniformly circular at the L4 and L5 level in newborns (Atilla et al. 1997), but over the first two decades of life a triangular configuration develops, caused by the differential changes in canal size in the sagittal and coronal planes described above. An archaeological study of 185 skeletons found a true trefoil configuration of the canal was not present in any skeleton until the age of 16 but was present in approximately 25% of adult skeletons at L5 with fairly consistent prevalence among all adult age groups (Papp et al. 1995). No clear association between the presence of a trefoil canal and the presence of degenerative changes was found and these findings are consistent with a trefoil canal being a developmental rather than congenital phenomenon.

1.4.2 Developmental lumbar spinal stenosis

As noted in §1.1, the term developmental lumbar stenosis was coined by Verbiest (1976) to describe failure of a normally formed canal to grow to normal adult sizes. This was contrasted with congenital stenosis, where malformation of the spine is present and associated with pathological canal narrowing, such as that occasionally seen in spina bifida (Sarpyener 1947). Developmental stenosis was thought to develop during childhood, only to become symptomatic in adulthood in a similar way to cervical ribs or the Arnold-Chiari malformation.

In his original case series of developmental LSS patients, Verbiest found the IPD on radiographs to be normal, and hence concluded that the narrowing must be in the AP plane. He performed an intraoperative measurement of the APD on a single patient and found it corresponded to the lower limit of previously published normal values (Verbiest 1954). After going on to perform *"a fair number of measurements"* on cadaveric skeletons and contrasting these measurements to a larger series of his patients, Verbiest defined the diagnostic thresholds listed below. The exact methodology by which these thresholds were reached is however unclear from the paper (Verbiest 1976).

- APD $<$ 10 mm — absolute stenosis
- APD $<$ 12 mm — relative stenosis

Verbiest believed that absolute stenosis was sufficient to cause symptoms alone, and hence always a surgical indication. Relative stenosis, he thought, implied a high risk of developing symptoms in the future.

Narrowing in the AP direction without changes to the IPD implies shortening of the pedicles in the presence of a otherwise normally sized vertebra. Consistent with this, Singh et al. (2005) compared 15 patients diagnosed with developmental LSS with 15 age matched controls and found shorter APD measurements in the developmental stenosis group were associated with smaller pedicle lengths, but no reduction in size of the vertebral body. Similarly, J. P.-Y. Cheung et al. (2014b) compared 100 patients requiring surgery for developmental LSS with a control group, and again found significantly smaller APD measurements in the stenosis group with no difference in IPD measurements or vertebral body size. Both these studies suffer from selection bias — patients within the stenosis groups had been chosen in part on the presence of anatomical changes their surgeons thought consistent with stenosis.

Other authors disagree that developmental stenosis is only manifested by changes in the APD — C. K. Lee et al. (1978) classified 3 types of stenosis: the *"sagittal flattening"* described by Verbiest, a concentric narrowing of the canal with short APD and IPD measurements, and developmental stenosis secondary to abnormal articular processes; either abnormally large or medial to their normal position. Evidence for the existence of this second group was found by Kitab et al. (2014), who in a study of 66 patients with symptomatic LSS presenting before the age of 50, found both the mean APD and IPD measurements were significantly smaller than a control group. This form of stenosis however appears rare; narrowing of the APD is twice as common in population studies as narrowing of the IPD (Eisenstein 1977).

Verbiest did not believe that degenerative LSS existed: he had never observed direct compression of the canal contents by osteophytosis during surgery, thought facet joint enlargement a purely developmental phenomenon, and did not believe the ligamentum flavum could compress the canal contents in any circumstance (Verbiest 1976). In a review published in 1976, he specifically cites Van Gelderen (1948), who provided the first description of neurogenic claudication in a patient with ligamentum flavum thickening, in order to dismiss this as a possibility. It has, however, become apparent with the use of modern imaging modalities that degenerative narrowing of the canal is by far more often the primary pathology. A study of participants

of the Framingham Cohort study showed measurements consistent with Verbiest's relative stenosis in only 4.7% of the population and absolute stenosis in 2.6%. In contrast, acquired degenerative stenosis was present in 22.5% (Kalichman et al. 2009). Furthermore, even in patients with measurements consistent with developmental LSS, the site of maximal narrowing generally does not occur at the site of maximal bony narrowing, but at the disc level. As has already been stated, measurements of the bony canal often poorly reflect the underlying true narrowing of the canal caused by the degenerate ligaments and disc (Bolender et al. 1985). Verbiest's failure to realise this may reflect his dependence in his studies upon operative findings and dissection; both techniques will often remove the posterior ligamentous structures to access the canal before measurements can be performed.

Narrower bony canals have been shown to increase the risk of disc herniations becoming symptomatic (R. Porter et al. 1978) and the risk for developing non-specific low back pain (Kalichman et al. 2009). It seems likely that developmental stenosis is not truly a separate condition to degenerative stenosis, outside perhaps a small percentage of individuals presenting well before onset of degeneration, but a risk factor for development of symptomatic degenerative LSS. Developmental narrowing reduces the degree of degenerative change required to compress the canal contents. In this narrative, developmental stenosis patients present at earlier ages to those with pure degenerative stenosis as they become symptomatic with degrees of degenerative change that would not trouble other individuals, rather than through direct compression of the canal contents by the developmental narrowing.

Considered as a risk factor rather than a disease in itself, radiological thresholds for diagnosing anatomical degenerative stenosis will always be to some extent arbitrary, as the risk for compression with future degenerative change will exist on a spectrum defined by the degree of developmental narrowing. Diagnostic thresholds may however have some clinical utility in allowing clinicians to judge the relative contribution of developmental changes to the any anatomical stenosis in a given patient. Numerous such thresholds, in addition to Verbiest's definitions, have been proposed in the literature and are listed in table 1.1. There are a number of significant limitations to the evidence presented. Many of the included studies use the lowest measurement or lowest 2.5th percentile within the group classified as normal as their diagnostic threshold. While this convention for defining normal is common in the medical literature there is no guarantee that the risk of developing symptomatic LSS is closely associated with this threshold. Other papers use comparison of measurements in a case and control group to select the threshold best separating the groups, but all such included papers either implicitly or explicitly define their case and control populations based upon the appearance of the canal on imaging, building a significant selection bias into the results.

The causes for developmental narrowing of the lumbar canal are not completely understood. There is evidence that symptomatic LSS has a strong genetic component (Battié et al. 2014), though the degree to which this represent a genetic predisposition to degeneration rather than a smaller bony canal is not known. Other hypotheses have included a role for childhood malnutrition or periods of significant physiological stress. Archaeological evidence from Native American populations undergoing a transition from a hunter-gather to an early agricultural lifestyle (associated with a relatively poor protein intake) found significantly smaller canals in those who grew up in the latter time period, and this effect persisted after controlling for age, gender and other cultural factors (Clark et al. 1985). The canals differed in mean APD, but showed no difference in IPD measurements. Similar archaeological studies have shown a relationship between dental manifestations of physiological stress and smaller bony canals—dental hypoplasia correlated with small IPD measurements and Harris lines correlated with small APD measurements (R. W. Porter et al. 1987). Narrowing of the APD is also more common in canals showing a trefoil morphology at L5 (Papp et al. 1995).

A role for an environmental insult during childhood may in part explain the preponderance of stenosis at the L4/5 and L5/S1 levels; alongside the natural hour-glass shape of the canal and propensity of degenerative change to affect the lower lumbar levels (discussed in the next section). As described in §1.4.1 maturation is slower at lower lumbar levels, with L4 and L5 only reaching mature sizes by the fourth or fifth years of life. This longer period of growth may increase the vulnerability of these levels to stenosis by expanding the window of opportunity for an environmental insult to occur and growth to be retarded.

While developmental stenosis as defined by Verbiest is rare, there is evidence that developmental narrowing of the canal as a whole is becoming more common. In a paper titled *"Secular changes of spinal canal dimensions in Western Switzerland: a narrowing epidemic"*, Schizas et al. (2014), in a study of CTs in 264 patients undertaken for abdominal pathology, found VF-CSA measurements significantly smaller in individuals borne between 1970 and 1979 compared to those borne between 1940 to 1949. This was contrary to what was expected, given the steady increase in population height with time in the West, and raises the possibility that symptomatic spinal stenosis as a whole may become more prevalent as younger generations age.

1.4.3 Segmental degeneration

The lumbar spine and its canal does not become anatomically static with completion of growth, but with age begins to accumulate changes presumed to be degenerative in nature. These changes can affect almost all parts of the lumbar spine and begin to be present just after adolescence in some individuals (Jensen et

Table 1.1: A table of proposed thresholds for developmental lumbar spinal stenosis (DLSS) developed from J. P.-Y. Cheung et al. 2014a

Citation	Population	Modality	Threshold (mm or mm^2)	Method
Verbiest 1979	116 Patients Cadaveric controls	XR	APD: Absolute stenosis < 10 Relative stenosis < 12	Method unclear, lower limit of normal compared to operative findings
Cheung 2014	100 DLSS undergoing surgery 100 asymptomatic controls Age 15–86, all Chinese	MRI	APD: L4 and L5 < 14 S1 < 12	ROC Analysis Prioritised sensitivity
Eisenstein 1977	443 Cadavers, South African 2166 complete vertebrae	Cadaveric XR	APD < 13 XR: APD < 15	Unclear
Chatha 2011	100 Patients with suspected metastasis Age 4–94, Canada	MRI	DS-APD: L1 < 9.3 L1/2 < 8.0 L2 < 8.8 L2/3 < 5.9 L3 < 7.7 L3/4 < 3.8 L4 < 7.2 L4/5 < 5.0 L5 < 6.9 L5/S1 < 5.5	2.5% centile cutoff
Lee 1995	90 Cadavers, Korean Age: 19–70 years	Cadaveric	APD < 11.8 XR: APD < 14	2.5% centile cutoff
Singh 2005	15 DLSS undergoing surgery 15 Asymptomatic Controls Age: 41–65 years, USA	XR MRI	Pedicle-Vertebra ratio: XR < 0.43 MR < 0.36 MR Pedical length < 6.5 MR Pedicular CSA < 213	Case-control study Manual cut-off selection
Ullrich 1980	60 abdominal CTs Age: 18–74	CT	Pedicular Axial images: APD < 11.5 IPD < 16 CSA < 145	Lowest limit of included patients

al. 1994); degeneration is often not associated with symptoms, casting doubt on a purely pathological aetiology and role. As the structures of the canal degenerate they tend to narrow the spinal canal, eventually causing degenerative anatomical lumbar spinal stenosis in a subset of individuals. The following sections describe these degenerative processes and the available evidence on how such changes are induced and propagate with age.

Intervertebral disc degeneration

The overall population of cells in the adult intervertebral disc is sparse and evidence from animal models suggests that as individual age a processes of progressive cell senescence and death occurs, depleting this population even further (Zhao et al. 2007). Cell type alterations also occur, with reductions in the number of proteoglycan producing cells derived from the fetal notochord and their replacement by chondrocytes (F. Wang et al. 2016). These changes are associated with increased inflammatory responses and enhanced catabolic metabolism (Clarençon et al. 2016), and overall result in a progressive reduction in proteoglycan content with age — down from 65% of the pulposus dry weight in young adults to about 30% dry weight at 60 years of age (Zhao et al. 2007).

Proteoglycans are responsible for retaining water within the disc, and consequently their loss results in the disc becoming dehydrated, observable as a progressive loss of signal on T2 weighted MRI studies (Pfirrmann et al. 2001). This compromises the force transmitting and spreading role of the pulposus, resulting in asymmetric loading of the end plates and direct load bearing by the annulus (Niosi et al. 2004). At later stages of degeneration the disc begins to loose height and the annulus bulges outwards, encroaching on the canal contents (Pfirrmann et al. 2001). The lax annulus may also allow abnormal translation of the vertebral bodies relative to each other, increasing mechanical loading on the facet joints and consequently altering spinal stability (Izzo et al. 2013b).

While, as described above, the degenerate disc may diffusely bulge beyond the limits of the vertebral bodies, more focal disc *herniations* may also occur, diagnostically separated from disc bulges by their involvement of less than 25% of the disc circumference (Fardon et al. 2014). Herniations are most commonly seen at the L4–5 and L5-S1 levels (approximately 90% of herniations) and may be superimposed upon more

diffuse disc bulges (Wilmink 2010c).

The term herniation is generally avoided in radiology reports in preference for sub-division of the pathology into either: protrusions, where the herniation has a broad based connection with the disc; extrusions, where the connection to the disc is narrower than the maximum extent of the herniation; and sequestrations, where the herniated disc material is no longer connected to the disc (Fardon et al. 2014). In protrusions the herniation remains fully covered by the annulus. In contrast, extrusions and sequestrations are thought to represent pulposus material that has been expressed beyond the confines of the annulus. These lesions therefore require a fissure in the annulus through which the pulposus can migrate, and such fissures may be demonstrated directly by discography, a now largely superseded imaging technique where a radiographic contrast agent is injected directly into the disc (Schellhas et al. 1996). Fissures can also be demonstrated as high intensity zones on T2 weighted MRI imaging and such findings are highly predictive of progressive degeneration at the same motion segment, leading some to suggest that formation of such fissures, perhaps through minor trauma, may be the initiating event in a cascade of degenerative processes (Sharma et al. 2011). Once pulposus material has extended through a fissure and escaped the annulus it is not completely free to move but is usually constrained by the boundaries of the anterior epidural space (Schellinger et al. 1990). Consequently they do not usually cross the midline due to the midline septum extending from the posterior longitudinal ligaments.

Disk herniations, while often considered to be a finding indicating pathology, can be present but asymptomatic in up to 76% of older individuals (Boos et al. 1995). There are several mechanisms which may underlie their variable association with symptom generation. One determining factor may be the ability of adjacent nerve roots to avoid compression (Wilmink 2010c), and indeed herniations with apparent nerve root compression on imaging have a much closer association with radicular pain than herniations alone (Boos et al. 1995). Nerve roots are constrained in their movement primarily by their attachment at the conus, the dorsal root ganglion and by the axilla and body of the nerve root sleeve. Nerve roots that have longer sleeves, i.e. those in the lower lumbar spine (Wilmink 2010c), may hence be more vulnerable. Similarly degeneration of adjacent structures or congenital narrowing of the canal at the site of the disc herniation may constrain the available space for the nerve root to escape compression. R. Porter et al. (1978) noticed that patients with symptomatic disc herniations tended to have smaller osseous canals with 68% of his symptomatic sample having canal APD below the 10th percentile of normal subjects. He also found a inverse relationship between the severity of symptoms and response to treatment with canal size. Similarly Kornberg et al. (1985) found that patients with symptomatic disc herniations severe enough to require surgery had significantly smaller bony canals than those who could be managed conservatively. A further feature which may be important in herniation related symptom generation is the type of herniation. The nucleus pulposus appears to have strong pro-inflammatory properties. McCarron et al. (1987) injected pulposus material into the epidural space of dogs and observed an intense subsequent inflammatory reaction involving the dural and nerve roots followed 14 to 21 days later by fibrotic scarring. Herniated pulposus material has proved to contain high levels of phospholipase A2, an enzyme known to be involved in inflammatory mediator production (Piperno et al. 1997). This inflammatory process may be responsible for the spontaneous resorption of herniated disc material that has been repeatedly reported—histological examination of sequestered discs has shown neovascularization and macrophage infiltration, suggesting an absorption process (Ito et al. 1996). The pro-inflammatory nature of the pulposus may underline the stronger association of disc extrusion the back pain compared to protrusions (where the annulus remains intact) — in a young asymptomatic cohort disc protrusions were seen in 27% of individuals but extrusions in 1%, whereas, in those with back pain extrusions were seen in 27% (Jensen et al. 1994).

In over half of cases, disc degeneration is associated with degenerative change in the vertebral end plates and associated marrow changes in the underlying vertebral body (de Roos et al. 1987). The end plates become thinned, forming fissures, fractures and clefts (Niosi et al. 2004) and eventually the defects may allow pulposus material to herniate through the endplate and into the adjacent vertebral body to form a Schmorl's node. The latter appear to be present with much higher frequency in patients with symptomatic LSS, and have been hypothesised to enhance the rate of degeneration presumably by reducing the volume of available pulposus material to distribute loads within the disc (Abbas et al. 2017).

The marrow changes underlying the degenerate end plate have been suggested to progress through a three stage process (oedema, fatty change and finally sclerosis) and are similar to those seen adjacent to other cartilaginous joints (Modic et al. 1988). The first of these stages (Type I Modic change — bone marrow oedema) has become somewhat controversial after suggestions that, unlike other grades of marrow change, it is directly associated with low back pain (Brinjikji et al. 2015a) and may be associated with low grade bacterial infection (Urquhart et al. 2015). An randomised controlled trial (RCT) found antibiotic treatment in the presence of Modic type 1 efficacious for improvement of the back pain and associated disability (Albert et al. 2013) but a more recent follow up RCT failed to replicate the benefit (Bråten et al. 2019). This has sparked a split between "believers" and "unbelievers" (Jensen, Ole K 2019) over the use of antibiotics in back pain.

Zygapophyseal joint and ligamentum flavum degeneration

The zygapophyseal joints are particularly vulnerability to degeneration due to their high level of mobility and large forces they must transmit (Clarençon et al. 2016). The degree of force transmitted through the joints appears correlated with disc degeneration — as the discs load bearing function is compromised the the facet joins may come to bear up to 70% of the axial load (Adams et al. 1983; Niosi et al. 2004). While disc degeneration may be seen without appreciable zygapophyseal joint degeneration, the converse was not observed in a consecutive series of 183 patients with LBP studied by Fujiwara et al. (1999), suggesting disc degeneration precedes, and potentially causes, zygapophyseal joint degeneration.

Zygapophyseal degeneration begins with formation of focal erosions at the margins of the joint cartilage, particularly at points of potential overload, such as the superior pole of the superior articular facet where the inferior facet causes maximal pressure during flexion (Kalichman et al. 2007). The focal erosions eventually coalesce into a diffuse loss of cartilage, leading to loss of joint space. Osteophytes begin to develop, preferentially at the anteromedial aspect of the joint, and associated hypertrophy of the facet joint is often the most predominant imaging feature (Grenier et al. 1987). This hypertrophy may however be an adaptive feature, hypothesised to help stabilise the joint (and associated motion segment) preventing unwanted translation movements of the vertebral bodies which may become possible due to disc pulposus loss and the associated slack disc annulus (Serhan et al. 2007). Rarely degenerate zygapophyseal joints may form synovial cysts which can cause compression of the canal contents (Doyle et al. 2004).

Alongside disc degeneration, other risk factors for zygapophyseal joint degeneration include advanced age, more caudal vertebral levels (presumably due to the increased loading and frequency of disc degeneration) and facet joint orientation (Kalichman et al. 2007). Jentzsch et al. (2013) found that a facet angle of greater than 32° to the coronal in the upper spine appeared to be independently associated with degeneration, but this study could not determine if the more coronal orientation was not in fact induced through remodeling of the joint surfaces by osteophytosis. Other studies have suggested the apparent orientation of the articular facets changes with age (J. Wang et al. 2009) suggesting this may be indeed be the case.

Associated with degeneration of the zygapophyseal joints, the ligamentum flavum may also undergo hypertrophy, and contribute to canal narrowing (Karavelioglu et al. 2016), though this role was initially controversial, with many believing that the apparent thickening of the ligament on imaging represented mere bucking of the ligament secondary to loss of disc height (Grenier et al. 1987). In a case series of 45 patients operated on for degenerate lumbar spinal stenosis, Yoshida et al. (1992) found that the thickening of the ligamentum flavum observed on pre-operative CT imaging in these patients was associated with clear histological changes: the normal elastin fibres had become irregular in arrangement and few in number, with associated fibroblast proliferation and a general increase in type II collagen fibres. These changes were most pronounced adjacent to the facet joint and were not present in samples from a younger control group who had been operated on for symptomatic acute disc herniation. Similar changes were seen by Schrader et al. (1999) who also found the changes were commonly associated with calcification of the ligament, and could be correlated to the patients age. Loss of elastic function of the ligamentum flavum and replacement by collagen fibres, may promote buckling of the stiffened ligament flavum into the canal during extension, rather than the normal elastic compression under load (Grenier et al. 1987).

Instability and Spondylolisthesis

A consensus definition of spinal instability is lacking (Muto et al. 2016b): Pope et al. (1985) defines instability as a loss of stiffness leading to abnormal and increased movements in the motion segments, while Izzo et al. (2013a) defines instability as the loss of the spine's ability under physiological loads to maintain its morphology such that no neurological deficit, deformity or incapacitating pain develops. Instability is frequently linked to other spinal conditions (Muto et al. 2016b). An initial pathological stimulus leads to pain and a protective muscular response, but one that leads to adoption of an abnormal posture, this changes the stresses placed on the various parts of the motion segment and promotes degeneration. Degeneration leads to further instability through pain, associated muscular responses, and direct alteration of the normal stabilising structures discussed in the previous section (Izzo et al. 2013b). This positive feedback cycle potentially allows degeneration to propagate between motion segments, but may eventually be balanced by a phase of re-stabilisation caused by facet joint hypertrophy, fibrosis of joint capsules, and radial expansive remodelling of the vertebral bodies (Izzo et al. 2013b).

Spondylolisthesis is defined as an anterior slipping of one vertebral body on the next vertebral body below it (Clarençon et al. 2016). Spondylolisthesis comes in two forms, degenerative and isthmic, the former being distinguished by the presence of an intact neural arch, and the latter (otherwise known as spondylolysis) being caused by defects in the pars interarticularis (Macnab 1950). The abnormal translation of vertebral bodies seen in spondylolisthesis implies the presence of instability.

Degenerative spondylolisthesis is most commonly seen at L4/5, affects around 4 to 14% of the population, and is more common in women (Cowley 2016). In a study of CTs in 118 patients over the age of 55 years Love et al. (1999) found that the zygapophyseal joints were more sagittal in orientation in

patients with spondylolisthesis compared to a control group, and this change in angulation appeared to be related to degenerate remodelling. Similar findings relating to articular facet angle were found by N. H. Kim et al. (1995) and Chaput et al. (2007) found that the presence of an increased fluid space between articular facets visualised in the recumbent MRI was highly predictive of a previously unseen spondylolisthesis at that level on standing radiographs. It therefore seems likely that a sagittal facet orientation secondary to degenerative remodeling, in combination with the erosion of the facet articular surfaces, facilitates translation of the vertebral bodies relative to each other (Beresford et al. 2010). The tendency of the lumbar vertebra to slip anteriorly is likely created by the lumbar lordosis and anterior tilt of the superior aspect of the sacrum (Bogduk 2012).

Increasing degenerative spondylolisthesis is correlated with general reduction in spinal canal volumes and its presence exaggerates the effect of postural changes including those caused by standing posture (Miao et al. 2013). The effect of spondylolisthesis is likely to be particularly important for the lateral recess, where the superior articular process translates forward and may catch the transiting nerve root between itself and the disc/vertebral body, and the neural exit foramen, which straddles the site of translation and hence becomes increasingly distorted as anterior slip occurs (Wilmink 2010c). In a large cross-sectional study of 938 participants Ishimoto et al. (2017) found that degenerative spondylolisthesis was significantly related to clinical LSS but not to nonspecific low-back pain.

Isthmic spondylolisthesis is caused by breaks in the pars interarticularis which disconnect the vertebral body from the neural arch and zygapophyseal joints and hence allow independent translation of the vertebral body unconstrained by the zygapophyseal joints (Clarençon et al. 2016). The aetiology behind the defects is not fully understood, with postulated mechanisms including failure of duplicate lateral mass ossification centres to fuse, fracture during birth or early childhood or stress fracture in later life (Wiltse 1962). Of note, the degree of vertebral translation may be much greater than in degenerative spondylolisthesis, though the central canal tends to be spared significant associated narrowing as the lamina do not translated forward (Wilmink 2010c).

1.4.4 Degenerative anatomical LSS

The discussion in the previous sections have implied a sequence of degeneration that occurs over time as individuals age — a process that can be referred to as progressive segmental degeneration (Kalff et al. 2013). The proposed sequence is as follows: disc degeneration occurs first, leading to loss of segmental height. The altered biomechanics then promote arthritis of the zygapophyseal joints which loose cartilage undergo changes in orientation through remodelling. These changes together promote spondylolisthesis and/or associated instability that promotes further degeneration and may cause it to propagate to adjacent levels (Bagley et al. 2019). Overtime, hypertrophy of the facet joints and fibrosis of ligaments may stabilise the canal, halting progression of degeneration at the newly stabilised segments (Izzo et al. 2013b). The fact that relatively advanced degeneration is often observed at one intervertebral level while adjacent levels may be relatively preserved argues for some precipitating event that is specific to a particular disc (K. M. C. Cheung et al. 2009). Longitudinal imaging studies of the spine have identified annular fissures as significant predictors of further disc degeneration (Sharma et al. 2011), suggesting minor trauma with associated tear formation may be the precipitating event.

Degenerative anatomical LSS occurs as an end consequence of progressive segmental degeneration, with the cumulative degenerative change described above contributing to reducing the size of the ligamentous portion of the lumbar canal at the affected levels (Wilmink 2010c). As the canal reduces in size, the available space for the canals contents reduces, CSF is effaced from the dural-sac, and both nerve roots (Schizas et al. 2010) and vascular structures (Manaka et al. 2003) become threatened with compression.

As discussed in § 1.3.5, the size of the canal even in normal individuals is dependent on posture and gravitational loading, with the canal tending to reduce in diameter on lumbar extension and upright positioning. Penning et al. (1987) in a study of 12 patients with NC, performed supine CT myelography using foam wedges placed on the scanner table to place the participants such that their lumbar spines assumed flexed or extended positions. They found marked facet hypertrophy led to narrowing of the lateral recesses in all the symptomatic patients and, on moving from lumbar flexion to extension, the lateral recesses became completely effaced with loss of any visible extradural fat around the transiting nerve roots. Similarly, Ozawa et al. (2012) used an axial loading devices during an MRI study to simulate the standing posture in patients with symptomatic LSS. They found that applying the axial load caused significant reductions in measured lumbar DS-CSAs, and this was more pronounced in those patients with degenerative lumbar spondylosis. Similar findings have been confirmed by other studies (Y. K. Kim et al. 2013), including one which found the reduction in DS-CSA on axial loading was related to the degree of disc degeneration, with more severe degeneration being associated with greater DS-CSA reductions except at at the most severe grade of disc degeneration (where the disc was completely collapsed) (Ahn et al. 2009). This latter finding may represent re-stabilisation and the fact that no further segmental height loss through disc compression was possible.

The postural compression of nerve roots through the mechanisms described above are generally thought to lay behind generation of NC symptoms (Wilmink 2010c), described in more detail in the next section.

Before moving on, it is important to note that not all canal narrowing is necessarily pathological — many of the changes above are almost uniformly exhibited by the older population (some degree of degenerative changes of the lumbar spine are present in nearly 100% of those over 60 — K. M. C. Cheung et al. (2009)). In fact, a systematic review published by Brinjikji et al. (2015b) of studies reporting MRI findings in completely asymptomatic individuals found disc degeneration rose from a prevalence of 37% in those 20 years of age to 96% at 80 years of age, alongside progressive facet joint arthrosis and development of spondylolisthesis in 23% by the age of 60. The fact that such changes are seen even in asymptomatic individuals strongly suggests they may be part of the normal biology of ageing and not necessarily pathological. In fact even severe degrees of anatomical spinal stenosis, which should presumably be associated with nerve root compression, can be seen in asymptomatic individuals, as can disc extrusions, even when substantially indenting the dural-sac (Boden et al. 1990). Such findings pose significant problems for the definition of anatomical lumbar spinal stenosis by preventing a clear boundary between normal and abnormal patients. Attempting to address this problem is central to the aims of this thesis.

1.5 A working definition

The consensus definition of degenerative symptomatic lumbar spinal stenosis, its clinical prevalence, management, and difficulties with defining anatomical degenerative LSS on imaging.

1.5.1 The NASS consensus definition of LSS

The North American Spinal Society has produced a consensus definition of symptomatic spinal stenosis which has two major requirements: anatomical narrowing of the lumbar spinal canal combined with a clinical syndrome of radiating buttock and/or leg pain and/or fatigue that is provoked by upright expercise and relieved by adoption of a posture where the lumbar spine is flexed (i.e. neurogenic claudication). Both the anatomical changes and neurogenic claudication must be present for the diagnosis to be made (Kreiner et al. 2013).

In order to decide if the canal of a patient presenting with NC is narrowed, the guidelines suggest the use of clinical imaging and "well-defined, articulated and validated criteria" but gives no information on what these criteria should be or how they should be validated (Kreiner et al. 2013).

While these guidelines are helpful in providing a relatively clear definition of the clinical features expected in the condition, the vague statements about anatomical narrowing limit their clinical utility and are largely due to a lack of an evidence based imaging definition of anatomical LSS. The various proposed definitions of degenerate anatomical stenosis used in the literature are discussed in § 1.5.4.

It should be noted that even the NASS definition of LSS associated symptoms are not universally accepted within the literature, with marked variation in the clinical findings held to be secondary to observed anatomical LSS, as shall be discussed further in § 2.1.

1.5.2 Clinical features of symptomatic LSS

The hallmark symptom of symptomatic LSS, neurogenic claudication (NC), has been described above.

Almost all symptomatic LSS patients have a degree of non-specific back pain (Katz 2008). Adding radiating pain from the back into the buttocks or leg has has a sensitivity of 88% but only a specificity of 34% for significant lumbar canal narrowing (Katz et al. 1995). Additional features which add specificity for a final clinical diagnosis of symptomatic LSS include: bilateral symptoms, leg pain being more severe than the back pain, pain being relieved by adoption of a stooped posture when walking and by sitting, and reports of associated urinary disturbance. An older age at presentation also adds specificity (de Schepper et al. 2013).

Differentiating NC from vascular claudication can be difficult but is generally achieved through its association with patient positioning (Katz 2008). While vascular claudication will be exacerbated by any exercise, patients with NC typically find exercise where the lumbar spine is flexed much easier, for instance walking up a hill or a flight of stairs may be relatively well tolerated, while descending a flight of stairs (which extends the lumbar spine) may become impossible with severe symptoms to the extent that some NC patients may adapt by walking down backwards. In contrast, the converse would be expected in claudication of vascular origin. Similarly, NC patients may find cycling does not provoke symptoms even at similar levels of exertion to what provokes their symptoms when walking. Both vascular claudication and neurogenic claudication symptoms are relieved by rest, but only in neurogenic claudication does sitting or lying substantially aid recovery (Cowley 2016).

Neurological examination at rest is often not revealing (Katz 2008). The Romberg test may be positive, and patients may show numbness in the perineal region and numbness and/or loss of proprioception in the feet. The patient may demonstrate a wide based gait and absent Achilles reflex (de Schepper et al. 2013). Motor and sensory findings are typically mild and weakness is uncommon (Katz et al. 1995).

Limitation of walking ability due to neurogenic claudication is a key cause of disability in patients

with symptomatic LSS (Fanuele et al. 2000) and individuals with LSS are known to have greater limitation in day to day walking performance when compared to those with non-specific low back pain (Tomkins-Lane et al. 2012). Patients subjective assessment of any improvement in their walking capacity is strongly correlated with post-treatment satisfaction (Yamashita et al. 2003).

It should be noted that many of the studies assessing the diagnostic accuracy of various symptoms and signs, included in the systematic review by de Schepper et al. (2013) and summarised above, use the final clinical diagnosis as their reference standard (where they have defined symptomatic LSS at all). This reference standard is therefore in part made based on the same clinical presentation assessed. It is therefore difficult to assess whether the symptoms reported to be associated with the diagnosis of symptomatic LSS by such studies are truly associated with the specific underlying pathophysiology, or simply reflections of the what the medical community have categorised as being part of the disease. This problem will be a recurrent theme in subsequent sections.

Related conditions of the cauda equina

For contrast, it is worth also considering two further conditions that are attributed to cauda equina compression.

Radiculopathy (often referred to as sciatica in the earlier literature) refers to pain associated with a single nerve root (Wilmink 2010d). This pain also radiates from the back into the lower leg and is typically described as sharp, burning and/or cramping in nature. It can be associated with an unpleasant dysesthesia, typically unilateral, and is frequently associated with paraesthesia, sensory loss or motor weakness in the distribution of the affected nerve root. Physical examination can provoke the pain through applying tension to the involved nerve root (through the straight leg raise). While the pathological distinctions between radiculopathy and symptomatic LSS are blurred, radiculopathy has a much closer link in the literature with individual nerve root compression by disc herniations. McCulloch (1977), for instance, defined symptomatic disc lesions as involving neurological symptoms restricted to that expected for a single nerve alongside evidence of disc protrusion.

In contrast, cauda equina syndrome is thought to be related to rapid onset generalised compression of a large segment of the cauda equina, typically by a large disc extrusion occluding the spinal canal. It is characterised by pain, loss of sphincter function, sexual dysfunction and saddle anaesthesia, and is generally treated as a surgical emergency requiring decompression (A. Gardner et al. 2011). In a similar way to that seen for symptomatic LSS, there is considerable variability as to how it is defined in the literature (Fraser et al. 2009).

It seems likely that canal pathology can result in multiple overlapping clinical syndromes and there are not clearly defined ties between one form of pathology and one particular syndrome (Wilmink 2010d).

1.5.3 Prevalence and natural history

The incidence of anatomical/symptomatic LSS is dependent on the definition used and unsurprisingly varies widely (Cowley 2016).

The Framingham study investigated the prevalence of anatomical LSS, using the AP canal dimensions defined by Verbiest to define stenosis (relative stenosis: APD < 12 mm, absolute stenosis < 10 mm). On CTs they measured on bone windows at the mid-vertebral body level to assess for developmental stenosis, finding a prevalence of 4.7% and 2.6% for relative and absolute stenosis respectively. They then measured on soft tissue windows at the disc level to assess for degenerative stenosis, finding a prevalence of 22.5% and 7.3% respectively. Degenerative stenosis increased in frequency with age, being most commonly present in the over 60s group. Developmental stenosis had no relationship with age, as expected.

A Japanese population cohort study of 1099 subjects, with a mean age of 66 years, investigated the prevalence of symptomatic lumbar spinal stenosis using the NASS definition (Ishimoto et al. 2012). All participants had MRIs and anatomical LSS was graded on a four point system: none, mild, moderate and severe (without further definition). The prevalence of anatomical LSS was much higher than that of symptomatic LSS: moderate to severe central narrowing was seen in 76.5% of participants while the prevalence of symptomatic LSS was only 9.3%. Again, a relationship with age was noted, with prevalence roughly reaching a plateau above 60 years of age.

There is an absence of reliable evidence relating to the natural history of LSS. Kreiner et al. (2013) in the NASS guidelines suggests that in those with mild to moderate symptoms the prognosis is favourable in one third to a half of patients (Minamide et al. 2013) but there was no evidence to define the natural history in those with severe symptoms. The guidelines also state that rapid neurological decline is rare.

Patients are on average within their mid-60s at the time of initial presentation, often have had symptoms for many months before presentation and typically have multiple other comorbidities, most commonly joint disorders, hypertension and heart disease (Cummins et al. 2006). Various environmental risk factors for disc degeneration have been suggested, including: smoking, obesity, diabetes mellitus, hypertension and physical activities such as driving and lifting weights (Cowley 2016), but in the same large Japanese study as that described above, only diabetes (alongside aging) was found to be a significant environmental risk factor (Teraguchi et al. 2017). There does appears to be a significant genetic component to developing significant degenerative change, with a large classical twin study estimating a heritability for anatomical LSS of 66.9% (Battié et al. 2014).

In the United States of America (USA) by the 90s symptomatic LSS had become the most common

reason for spinal surgery in the over 65s (Ciol et al. 1996), and by 2007 greater than 37,000 operations were been performed annually on medicare patients, with a total cost of $1.65 billion per annum (Bagley et al. 2019). In the United Kingdom of Great Britain and Northern Ireland (UK), all forms of back pain cost the UK economy up to $12 billion per annum (Bibby et al. 2001).

1.5.4 Diagnostic testing in low back disorders

In order to make the diagnosis of symptomatic LSS in a patient with the correct clinical presentation it is necessary to confirm the presence of anatomical LSS. For this, clinical imaging is required, and such imaging can also help target surgical intervention and rule out dangerous differential diagnosis such as metastatic disease (Andreisek et al. 2011). It is important to realise that the presence of significant stenosis in asymptomatic individuals precludes the diagnosis of symptomatic spinal stenosis purely on anatomical grounds (§ 1.4.4), at least using simple measurements of the canal.

A systematic review by de Schepper et al. (2013) found MRI to be the most sensitive and specific test for the anatomical changes of spinal stenosis with the included studies usually using intra-operative findings as the reference standard to which the imaging findings were compared. The included papers found sensitivities ranging between 60% and 100% and specificities ranging between 43% and 99%, with the large ranges likely explained by the significant heterogeneity in methodology. MRI is recommended by the NASS as the first line investigation in suspected spinal stenosis (Kreiner et al. 2013), with more invasive CT myelography (CT after intra-thecal injection of a contrast adjacent to opacify the CSF within the dural-sac) considered the second line, to be used in those with contraindications to MRI imaging.

Other imaging modalities have largely been superseded by the above techniques (Wilmink 2010a), though plain radiographs retain a role in assessment of spinal curvature and alignment when standing, and similarly traditional myelography allows dynamic assessment of the dural-sac with varying patient positioning in a way not easily achievable with standard CT or MRI. As discussed in previous sections of the introduction, the use of axial loaded and standing cross-sectional imaging is now becoming more common, (e.g. Splendiani et al. 2014), but remains largely unavailable in the National Health Service (NHS) and is considered only a potentially useful adjunct to standard imaging by the NASS guidelines (Kreiner et al. 2013).

Despite the importance of the anatomical changes in the diagnosis of spinal stenosis, there is a little agreement on both on how these change should be measured and what level of change contributes stenosis. A systematic review of papers reporting radiological criteria to define LSS, found a large number of separate quantitative measurements used to define stenosis, with multiple different thresholds being used for most measurement (Steurer et al. 2011) — see tables 1.2 and 1.3. Across all the included papers, the most common quantitative definitions of stenosis were a DS-CSA of less than $100\ mm^2$, and LRD and NFD of less than $3\ mm$.

The rational behind the choice of these thresholds within the included papers was often unclear, but for cross-sectional measurements of the dural-sac, reference is often made to attempts published in the 1980s to derive a threshold definition of central LSS based upon a series of cadaveric studies. Schönström et al. (1984) studied the effect on intradural pressure caused by circumferential constriction of freshly dissected lumbar dural-sacs. He found pressure began to rise after constriction beyond a cross sectional area of $76.9\ mm^2$ at L2 falling to $64.8\ mm^2$ by L4 as the volume of neural tissue in the canal reduces. These thresholds presumably represent the cross-sectional areas at which all CSF has been effaced from the dural-sac leaving on the neural and vascular tissue. A similar follow up study in 1988 suggested the pressure increases occurred fairly consistently around $75\ mm^2$ across multiple different subjects (Schönström et al. 1988). Schönström went on to suggest a DS-CSA threshold of $75\ mm^2$ represented "absolute stenosis", and suggested a threshold of $100\ mm^2$ as "relative stenosis", after assessing the minimum size of the dural-sac on CT studies of 24 patients who had subsequently underwent surgery for symptomatic LSS (Schonstrom et al. 1985).

There are however numerous weakness with thresholds based on quantitative measurements including a relative lack of information on what should constitute normal, alongside if and how the thresholds should account for spinal level, changes in patient body size and demographics (Andreisek et al. 2011). There are also difficulties posed by the the effect of slice angulation (Henderson et al. 2012) and, as will be explored in detail in Chapter 2, the clinical validity of the quantitative measurements of the canal and dural-sac is far from clear, with little relationship between the various measurements and the presence of NC or symptom severity.

Perhaps because of these limitations and the large number of proposed thresholds to choose from, it should be no surprise that a Delphi survey of practising clinicians was unable to find a consensus on a single quantitative threshold for defining anatomical LSS (Mamisch et al. 2012). Instead, the clinicians had a clear preference for qualitative parameters, with the presence of causative degenerative change, effacement of perineural fat within the neural exit foramina, absence of CSF around the cauda equina and redundant nerve roots above the level of stenosis being cited as the most important factors in deciding upon the presence of anatomical LSS. There was however insufficient agreement to propose a consensus definition based on these parameters and a separate systematic review published by Andreisek et al. (2013) found 14 qualitative or semi-quantitative competing grading systems to define and characterise anatomical LSS.

Table 1.2: Various quantitative thresholds for central spinal stenosis on cross-sectional imaging used within the literature. Adapted from Steurer et al. (2011, table 1).

Measurement	Imaging Modality	Citation	Threshold
APD	MRI	Fukusaki et al. 1998	$< 15\ mm$
		Koc et al. 2009	$< 12\ mm$
	CT	Bolender et al. 1985	$< 13\ mm$
		Haig et al. 2007	$\leq 11.95\ mm$
		B. C. Lee et al. 1978	$< 15\ mm$ (Suggestive)
			$< 10\ mm$ (Diagnostic)
		Jönsson et al. 1997	$< 10\ mm$
DS-CSA	MRI	Hamanishi et al. 1994	$< 100\ mm^2$
		Mariconda et al. 2002	$< 130\ mm^2$
	CT	Bolender et al. 1985	$100 - 130\ mm^2$ (Early stenosis)
			$< 100\ mm^2$ (Present stenosis)
		Schonstrom et al. 1985	$< 100\ mm^2$
		Schönström et al. 2001	$100 - 130\ mm^2$ (Moderate stenosis)
			$< 75\ mm^2$ (Severe stenosis)
		Ullrich et al. 1980	$< 140\ mm2$
Stenosis Ratio	MRI	Laurencin et al. 1999	L3/L4: < 0.66
			L4/L5: < 0.62
			L5/S1: < 0.73
ILD	MRI	Herzog et al. 1991	L2/L3: $< 10\ mm$
			L3/L4: $< 10\ mm$
			L4/L5: $< 12\ mm$
			L5/S1: $< 13\ mm$
	CT	Wilmink et al. 1988	$< 11\ mm$
IFD	MRI	Koc et al. 2009	$< 15\ mm$
		Ullrich et al. 1980	$< 16\ mm$

In an attempt to bring some order to both radiology reports and subsequent research, Andreisek in 2014 published a set of core radiological criteria for use in description of anatomical LSS (Andreisek et al. 2014). These measurements and grading systems were picked primarily based upon on reported intra- and inter-observer reliability alongside perception of their practicality of use, rather than on any criteria of clinical validity, due to a lack of available high-quality evidence. This list of grading systems was used as a starting point when deciding what measurements of stenosis should be explored within the observation study presented in Chapters 3 – 8.

Core Reporting Criteria in Spinal Stenosis — Andreisek et al. (2014):

- Central Stenosis:
  - Compromise of the central zone — "Lurie" grading (Lurie et al. 2008b)
  - Relation between fluid and cauda equina — "Schizas" Grading (Schizas et al. 2010)
- Lateral stenosis:
  - Nerve root compression in the lateral recess — "Bartynski" grading (Bartynski et al. 2003)
- Foraminal Stenosis:
  - Foraminal nerve root impingement — "Pfirrmann" grading (Pfirrmann et al. 2004)
  - Compromise of the foraminal zone — "Lurie" Grading (Lurie et al. 2008a)
- Research Only — Quantitative measures:
  - Cross-sectional area of the dural sac
  - Lateral recess height

1.6 The pathophysiology of neurogenic claudication

An exploration of what may cause the key symptoms associated with symptomatic spinal stenosis

1.6.1 Postural compression alone does not explain neurogenic claudication

The larger part of the introduction above contains a clear narrative around symptomatic lumbar spinal stenosis. In this narrative, all individuals undergo degenerative change of the lumbar spine to some degree as they age and this degeneration leads to progressive narrowing of the spinal canal. Some individuals have a developmentally smaller lumbar canals (§ 1.4.2) and some individuals have a predisposition, genetic and/or environmental, for degeneration

Table 1.3: Various quantitative thresholds used for lateral recess and neural exit foramen stenosis within the literature. All listed studies used CT imaging. Adapted from Steurer et al. (2011, table 2).

Measurement	Citation	Threshold
LRD	Ciric et al. 1980	$5\ mm$ (Normal) $< 3\ mm$ (Highly indicative) $< 2\ mm$ (Diagnostic)
	Dincer et al. 1991	$< 5\ mm$
	Mikhael et al. 1981	$> 5\ mm$ (Normal) $3 - 5\ mm$ (Suggestive) $\leq 3\ mm$ (Definitive)
	Strojnik 2001	$\leq 3.6\ mm$
LRA	Strojnik 2001	$< 30^{\circ}$
NFD	Beers et al. 1985	$< 3\ mm$

(§ 1.4.3). Such individuals eventually develop a critical level of canal narrowing, usually occurring in their 60s or 70s, in which the cauda equina nerve roots starts to be compressed in a posture dependent manner. As the canal undergoes further narrowing the more it will be functionally narrowed by additional extension or axial loading — the rule of progressive narrowing proposed by Penning (1992).

Takahashi et al. (1995), using an intra-thecal pressure transducer placed by lumbar puncture, demonstrated significantly higher CSF pressures during walking in patients with spinal stenosis compared to a control group. This raise in pressure reflects nerve root compression, which leads to claudication symptoms, and associated nerve root dysfunction leads to lower limb weakness and autonomic disturbances. Surgical relief of the causative narrowing appears to provide symptomatic relief, presumably by preventing further compression, at least in the short term (Kalff et al. 2013).

This is the "textbook" explanation of neurogenic claudication which has become a staple of the secondary literature (e.g. Katz 2008) and is closely related to the mechanisms proposed in the original papers published by Verbiest 1976. There are however multiple points where this narrative breaks down. As already noted, a significant proportion of participants with significant degenerative spinal stenosis are asymptomatic (Ishimoto et al. 2013). In addition, R. W. Porter (1996) pointed out that patients with spinal stenosis symptoms are also likely to have had anatomical stenosis for many years before their presentation, given the slowly progressive nature of segmental degeneration and severe central stenosis can be seen in spinal tumours or large disc protrusions without claudication symptoms (R. W. Porter 1996). Nor is it apparent from this narrative why one form of generalised cauda equina compression should lead to cauda equina syndrome and another to neurogenic claudication, syndromes with incompletely overlapping characteristics.

The above argues that some additional pathological step is a necessity beyond simple canal stenosis for the generation of claudication symptoms.

1.6.2 Potential mechanisms of radicular pain

To reduce complexity, it is perhaps beneficial to step away from pathology affecting the whole cauda equina, and instead consider the pathophysiology of single nerve root compression in radiculopathy type presentations.

Confirmed compression of individual nerve roots by a localised disc protrusion or extrusion, while strongly associated with distal leg pain, can result in a wide range of degrees of pain and dysfunction that do not well correlate with the degree of observed compression (Beattie et al. 2000). For instance, Garfin et al. (1991) reported on four patients whose radiculopathy type pain had been ascribed to to observable nerve root compression by disc herniations but which gradual resolved without demonstratable change in the compressed nerve roots.

As was pointed out by Cavanaugh (1995), acute compression of a peripheral nerve does not result in pain, but rather loss of action potential conduction and hence nerve function, as is apparent to anyone who has awoken from sleep with a dead arm. Several intra-operative stimulation experiments seem to confirm a requirement for a process of nerve root sensitisation before compressive pain generation is possible. Smyth et al. (1958) during an operation for radicular symptoms, passed nylon loops around a nerve root that had been compressed by a disc protrusion and several adjacent uncompressed nerve roots. With the patient awake, traction on the nylon loop attached to the previously compressed nerve root led to severe radicular pain, but traction on previously uncompressed nerve roots did not result in the same degree of symptoms. Greenbarg et al. (1988) found that application of a local anaesthetic to the previously compressed nerve root could abolish the induced pain. Kuslich et al. (1991) found that both mechanical or electrical intra-operative stimulation of a compressed nerve root would replicate the patients symptoms, while previously uncompressed nerve roots were insensitive to such stimuli.

Pain in the presence of a normal nervous system requires the stimulation of nociceptors. In contrast, direct stimulation of action potentials within injured nerve root fibres is thought possible and to

lie behind de-afferation type pains (Wilmink 2010d). Both these mechanisms of pain generation may be involved with the pain associated with sensitised nerve roots, with the latter presumably generating the radicular pain experienced in the segmental area supplied by the injured nerve root, and the former hypothesised to contribute to "pseudo-radicular" pains poorly corresponding to the associated nerve roots supply (Wilmink 2010d). This leaves the question of by what processes is a nerve root sensitised and why it affects some compressed nerve roots and not others.

Chronic neuronal injury is known to result in histological changes including microvascular changes, inflammatory infiltration and oedema of the nerve with associated demyelination of axons. This process appears to manifest days to months after the initial injury (Rydevik et al. 1984). Compressed and symptomatic nerve roots show prolonged contrast enhancement on MRI imaging after administration of gadolinium, arguing for a either for a breakdown in the blood-nerve barrier or development of neovascularity (Jinkins 1993). Similar features have been demonstrated for the cauda equina at the site of compression in spinal stenosis.(Kobayashi et al. 2015).

There is also a potential role for inflammation generated by extruded nucleus pulposus material. As already noted in § 1.4.3, herniated pulposus material contains pro-inflammatory enzymes (Piperno et al. 1997) and generate intense inflammatory reactions when introduced into the extradural space during in vivo animal studies (McCarron et al. 1987). The same inflammatory mediators, when introduced directly into the lumbar epidural space of rats induced demyelination in exposed nerve roots, with an increased propensity to ectopic discharge on compression (Chen et al. 1997). This does not however appear to to be the entire story, Omarker et al. (1998) found the presence of herniated material from a punctured disc alone was insufficient to induce significant changes in thresholds for nerve root activation using either compressive or thermal stimuli, but herniated material with associated nerve root displacement did.

The above argues that generation of pain from single nerve root compression within the lumbar canal is likely to be a multi-factorial pathway with contribution from direct compression, inflammatory processes and histological changes in the nerve roots secondary to chronic injury.

1.6.3 The multi-level compression hypothesis

R. W. Porter (1996) suggested a hypothesis for symptom production in symptomatic LSS based upon two level compression. He noted that the veins of the nerve roots do not anastomose and generally drain distally to the neural exit foramen or, if this is occluded, proximally to the conus. Hence a two level compression would be required to cause venous congestion and associated oedema, potentially explaining why not everyone with apparent stenosis has symptoms.

The capillaries and venules of the nerve roots have been shown to be occluded by relatively low pressures (approximately 30 to 40 mmHg, Olmarker et al. 1989), well below the elevated pressures found in the stenotic canal of patients with neurogenic claudication, where during extension pressures of 80 to 100 mmHg can be reached (Takahashi et al. 1995).

Porcine cauda equina models have shown a single low-level compression causes little effect on function, but a two level compression at low pressures causes significant impairment of nerve conduction with failure of the normal arterial vasodilation in response to repeated electrical stimulation (Olmarker et al. 1992). Similarly, rat models have shown venous stasis can result in ectopic firing and nerve dysfunction (Ikawa et al. 2005) and examination of dog cauda equina which had been exposed to either arterial or venous occlusion at the level of the aorta or IVC showed far greater evidence of blood-nerve barrier breakdown in the case of venous occlusion compared to arterial ischaemia (Kobayashi et al. 2008).

Direct observation of the vessels surrounding the cauda equina by myeloscopy in patients during episodes of neurogenic claudication while walking show changes consistent with venous stagnation, and this does not occur in control patients under the same provocation (Ooi et al. 1990). During lumbar extension in non-stenotic individuals, CSF pressure rises are thought to be avoided by displacement of blood within the valveless and compressible anterior epidural plexus (Penning et al. 1981). In individuals with NC this plexus has been shown to be significant effaced compared to a control group, with multilevel involvement showing a correlation with greater symptoms (Manaka et al. 2003). It is possible multilevel compression of this plexus reduces its ability to drain under pressure, potentially reducing the ability of the dural-sac to bulge during postural changes and prompting CSF pressure rises.

If Porter's hypothesis is correct, we would expect to see greater incidence of NC symptoms in individuals with demonstrable multi-level compression and potentially a correlation between number of stenotic levels and symptoms — evidence around this subject is collected in subsequent chapters.

1.6.4 Potential contributions to pain from other structures

Other structures in the lumbar spine aside from the nerve roots may also contribute to pain generation. Both the disc annulus and facet joints are known to be richly innervated (S. Roberts et al. 1995; Kalichman et al. 2007) Some of the same intra-operative studies which explored sensitisation of nerve roots to compressive stimuli cited in § 1.6.2 also found other structures appeared sensitive to stimulation. Kuslich et al. (1991) for instance found stimulation of the annulus fibrosis reproduced a patient's low back pain in about two thirds of patients. As previously discussed, Modic I end plate change also appears to be re-

lated to low back pain (Brinjikji et al. 2015a), and zygapophyseal joint arthrosis is thought to account for 14% to 40% of low back pain directly, has been studied by direct stimulation of the joints nerve supply, and the pain associated with such structures may also appears "pseudo-radicular" in nature, radiating into the buttocks, groin and thighs and potentially mimicking neurogenic claudication (Beresford et al. 2010).

From the discussion above, the pathogenesis of the clinical features in symptomatic spinal stenosis seems therefore likely to be multifactorial, with contribution from multi-level compression and venous congestion, direct compression effects with inflammatory sensitisation of the nerve roots and potentially referred pain from other sources (Muto et al. 2016a).

1.7 An agenda

Summary so far, common methodological problems relating to the lack of a reference standard for LSS, and an initial set of research questions…

1.7.1 Problems caused by the lack of an accepted reference standard

In the previous section the argument was made that some additional pathological step, beyond compression alone, was required for generation of the claudication symptoms, or alternatively that claudication was generated in part through painful degenerative change affecting other spinal structures manifesting in a posturally dependent manner. In the most extreme form of this latter explanation, the observed narrowing of the canal in patients with neurogenic claudication is simply a confounding variable, coexisting with the actual cause of the patients symptoms.

While it is clear the standard "textbook" explanation of the pathophysiology of symptomatic LSS is too simple, it is not clear from the literature which or to what degree the potential alternative explanations contribute to the symptom generation in symptomatic LSS. These ambiguities also generate difficulties in creating a useful definition of anatomical LSS — no simple definition of anatomical stenosis will capture all symptomatic patients or exclude all non-symptomatic patients given the previously discussed disconnect between anatomy and symptoms. This ambiguity then manifests in the literature by variability in definitions of anatomical stenosis between papers, lack of agreement by clinicians on what measurements of the canal are relevant, and, perhaps unsurprisingly, the failure of the consensus NASS definition to provide any useful guidance as to what degree of anatomical narrowing should be considered pathological, as discussed in § 1.5.1 and § 1.5.4.

The attempt to find an evidence based definition of degenerative anatomical stenosis within the literature has generally fallen into a set of approaches, all of which are associated with significant issues.

Some attempt to derive an anatomical definition from the presumed pathological mechanism — for instance Schönström's 1984 measurements of the the cross-sectional area of the dural sac in cadavers. In order to produce useful results, such methodologies requires the underlying assumptions regarding the presumed pathological mechanism to be correct, and also the experimental measurement to be reliably transferable from the *in vivo* measurement to measurement with imaging.

Some use a degree of deviation from a "normal population". This approach is common in the papers investigating developmental stenosis, as seen in § 1.4.2, and often takes the form of declaring the bottom or top 2.5% of a population for a given measurement abnormal. Such thresholds are however ultimately arbitrary and therefore do not necessarily correspond to disease status, risk of developing disease or have useful prognostic implications. There are also significant difficulties in choosing an appropriate "normal population" who are then measured to generate the threshold measurement defined as abnormal. As already discussed, the canal even in asymptomatic populations changes with age, gender and ethnicity and hence any reference canal used should require similar age, gender and ethnicity matching. While a healthy asymptomatic population is perhaps the most natural population to choose, this also can cause problems as almost all individuals will experience episodes of low-back pain at some point in their life (Wong et al. 2017), and individuals with longstanding low back pain are known to have on average smaller canals than those who report no recent episode of back pain (Hamanishi et al. 1994). As patients with NC often also have non specific low-back pain in addition to the claudication symptoms there is a risk that comparison to individuals without any low-back will exaggerate any difference in canal size measurements between groups, exaggerating the sensitivity and specificity of the measurement.

Some compare the performance of some imaging based measurement in dividing diseased from non-diseased individuals to the performance of a reference standard for the same (for instance the studies reviewed in de Schepper et al. 2013). In the case of papers investigating symptomatic LSS, the reference used is often findings on surgical exploration. While this methodology is well established and can be made rigorous, it ultimately assumes that the reference standard is providing useful information about the disease status of the patient. In the case of symptomatic LSS, the reference standards own evidence base is often poor or absent (Andreisek et al.

2011) and the justification for its categorisation of patients is usually based on reasoning from the purported pathological mechanism of symptom generation — i.e. compression of the cauda equina. Showing that MRI demonstrated severe anatomical stenosis has high sensitivity and specificity for finding cauda equina nerve root compression on surgical exploration or myelography only demonstrates that the MRI accurately depicts the anatomy but does not confirm that MRI accurately depicts disease status.

Some deal with the absence of a reliable reference standard by replacing it with a consensus diagnosis. For instance, Haig et al. (2006b) created such a consensus diagnosis using a decision by an expert panel who considered clinical findings, imaging and other diagnostic test results. The sensitivity and specificity of the tests of interest for dividing patients into diseased and non-diseased groups as defined by the consensus diagnosis was then found. This methodology has advantages, particularly in the evaluation of some less expensive or invasive investigation that could potentially replace the investigations undertaken in forming the consensus diagnosis. However, in the case of conditions where the consensus diagnosis is partly based on the result of some diagnostic test itself of uncertain validity, the consensus diagnosis will necessarily reflect its results and any underlying assumptions about the meaning of those test results. For symptomatic LSS the consensus diagnosis is usually based in part upon expert interpretation of anatomical imaging of the lumbar spine, as per the NASS definition of symptomatic LSS. Hence, when investigating the agreement between some new imaging based sign or measurement and the consensus diagnosis, it then becomes unclear whether the results truly represent an association with the true underlying disease status of the patient or simply reflects the preexisting anatomical requirements imposed by the imaging input in forming the consensus diagnosis.

A pragmatic approach aiming to overcome these difficulties may attempt to find a definition of anatomical stenosis based on the inclusion criteria of RCT or cohort studies. Such a definition, however arbitrary chosen by the initial studies, would be potentially be predictive of any treatment effect or prognosis established by the study in those individuals meeting the other inclusion and exclusion criteria. There is however no guarantee that a true causal association exists between any anatomical definition of stenosis used in any trial and a demonstrated treatment effect unless an analysis comparing treatment effect with varying anatomical definitions has been performed while other variables were held constant. In the case of symptomatic lumbar spinal stenosis Steurer et al. (2011) did not find any consistent quantitative definition of LSS within inclusion criteria of RCTs.

These difficulties are a recurrent theme within the literature, explain the lack of an accepted definition of anatomical lumbar spinal stenosis, and inform the research questions set out below.

1.7.2 Research questions

The overall goal of this thesis is to make progress towards a better evidence based definition of anatomical lumbar spinal stenosis, in the hope of aiding radiologists and clinicians make better, more consistent assessments of lumbar spine MRI imaging in suspected symptomatic lumbar spinal stenosis. To this aim, the following questions are posed, with the hope that answering them will advance this cause.

What is the best way to measure narrowing of the spinal canal on MRI?

As was reviewed in § 1.3.3 and § 1.5.4, there are numerous competing ways to measure the spinal canal. In attempting to select between measurements it is perhaps helpful to review some of the features an ideal anatomical diagnostic measurement should have. Note these features overlap and are interrelated.

Reliability: For obvious reasons, the measurement must have an acceptable inter- and intra-rater reliability. As part of this, the measurement should be resistant to changing imaging parameters, e.g slice angulation.

Completeness: the measurement should capture as much anatomical information as possible about the structure it attempts to measure. For instance, measurement of the dural-sac in a single plane (e.g. the APD) would not necessarily capture compression of the dural-sac in the transverse plane, whereas measurement of the cross-sectional area of the sac would.

Convenience: the measurement should be easy to perform quickly with as few steps as possible required by the measurement process. Clinicians have limited time and complex measurements requiring numerous steps are unlikely to gain wide-spread usage or, if used, to be used correctly, leading to problems with reliability. Similarly it should be measurable on standard imaging sequences/protocols to prevent the need for increased scanner time per patient, which is also a relatively scarce resource even in nations with developed health care systems.

Well defined normal range: the measurement should have a normal range that is clearly defined, at least on the side of the distribution on which pathology is thought to occur, with a good understanding of how the normal rage changes with demographic variables. The definition of normal must also be defined in a manner relevant to the condition the measurement is designed to detect, rather than a simply based on the general population or based on those who are generally asymptomatic without further consideration.

Diagnostic utility: the measurement should either clearly separate diseased individuals from non-diseased controls or provide enough discriminating power to combine usefully with other diagnostic tests to perform this function.

Prognostic implications: the greater the association of the measurement with the prognosis of the patient, with or without treatment, the greater its potential to

inform management decisions

From our review of the literature so far, no measurement clearly excels in the requirements listed above. Andreisek's consensus core reporting items were selected pragmatically based upon their reliability of measurement and coverage of various anatomical factors potentially related to the pathophysiology of neurogenic claudication (Andreisek et al. 2014). There is a clear preference both in the consensus criteria and Delphi survey results of practising clinicians for qualitative scoring systems (Mamisch et al. 2012), however this seems to represent an understandable bias towards scoring systems designed around the presumed pathophysiology of the condition — most included systems attempt to directly measure nerve root compression — and the clear problems with interpretation of quantitative measurements. Qualitative systems suffer from more problems with subjectivity than quantitative measurements and will fail if the underlying assumptions implicit in their design are not relevant.

Further evaluation of the performance of the available imaging based measurements of the lumbar spinal canal, both quantitative and qualitative, with relevance to the properties listed above is therefore warranted.

Is there a relationship between the spinal canal size on MRI and severity of clinical presentation in symptomatic LSS and, if so, what measurement best represents this association?

Given the putative pathophysiology in symptomatic LSS, a relationship between canal size and severity of clinical symptoms would be expected. More severe canal narrowing should result in a reduction of the degree of dynamic posturally dependent narrowing required to induce cauda equina compression, presumably causing symptoms to occur earlier during upright exercise. More severe narrowing would also be expected to reduce the effectiveness of any mitigation strategy the patient may adopt to alleviate the symptoms — e.g. by adoption of a more flexed lumbar posture. Measurements therefore with a demonstrable close association with clinical severity would presumably be measuring anatomical changes linked to the generation of symptoms and hence would be potentially be important in any definition of lumbar spinal stenosis.

A strong relationship between a given anatomical measurement and severity of clinical presentation is also potentially clinically useful even if that measurement is not suitable for use in a definition of anatomical stenosis. Older patients often have multiple comorbidities and a given clinical presentation may reflect contributing pathology from multiple different condition. For instance an older patient with diabetes could present with a mixture of symptoms from spinal stenosis, polyneuropathy and vascular claudication. A measurement of the spinal canal with a close relationship with severity of clinical presentation would allow judgement of how much a complex patient's symptoms are likely due to the spinal canal narrowing rather than any other pathology and hence facilitate decisions regarding spinal intervention in such patients.

What radiological features are associated with neurogenic claudication?

Given the lack of an accepted diagnostic reference standard, problems with implicit bias when trying to derive a definition of anatomical LSS by comparing symptomatic LSS patients to a control group, and reasonable doubts around the pathophysiology through which symptoms are generated in symptomatic LSS, a new approach to finding an evidence based definition of anatomical LSS is required.

One potential way forward is to start again from first principles, directly comparing imaging in patients with and without neurogenic claudication without specifying any anatomical inclusion or exclusion criteria in the case or control group definitions and hence avoiding any bias introduced by expectations of what the canal in symptomatic LSS should look like.

This approach has the potential to find anatomical features or a combination of anatomical features consistently associated with the key clinical feature of symptomatic LSS.

While the utility of such a measurement may not be immediately apparent — after all you don't need a test to tell you whether a patient has claudication or not, you can just ask them — such an anatomical feature would act as a useful starting point for forming a definition of anatomical stenosis and would potentially have pathophysiological significance.

Chapter 2

Literature Review

2.1 Scoping review

As the first step in addressing the research questions posed in § 1.7.2, a scoping review of observational studies in symptomatic lumbar spinal stenosis was conducted, aiming to identify and include studies investigating the relationship between MRI based measurements of the lumbar spinal canal with the clinical status of the patient. Scoping reviews are a form of rapid literature review, designed to be performed before embarking upon a full systematic review, in order to allow the review question, search strategy and planned methodology to be refined.

In addition, conducting a scoping review at the beginning of the PhD process also offered an opportunity to systematically build and assess a library of immediately relevant research. From each citation, data on methodology (including case and control group definitions, MRI measurements and clinical scoring systems used) and results relevant to the link between MRI findings and severity of clinical presentation and prognosis was extracted.

The full scoping review methodology is presented in Appendix B.1.1 and a summary of the results of the search strategy is presented in figure 2.1. Thirty-six citations were included in the final review.

2.1.1 Defining the case and control populations within the literature

The clinical definition of lumbar spinal stenosis within the included papers was variable (see figure 2.3). Only one paper directly referred to the NASS definition (Ishimoto et al. 2013). Neurogenic claudication was a required part of the diagnosis in only 33% of papers and was one of several qualifying features in a further 25% of papers. Many papers failed to give a clinical definition of symptomatic LSS as part of their inclusion criteria. In total, 78% of papers required radiological stenosis in their inclusion criteria for symptomatic participants, either implicitly or explicitly, however, most did not give a specific threshold for inclusion.

A variety of different sources for symptomatic LSS participant recruitment were used (see table 2.1). Participants were most commonly recruited from patients undergoing an intervention in a secondary care setting. Of the papers including a control group, the most common sources were asymptomatic volunteers and those with non-specific low back pain (see figure 2.2). Twenty-two different clinical scoring systems for severity of patient symptoms, disability and quality of life were used of which the Oswestry disability index (ODI) was the most common followed by visual analogue scale (VAS) based assessment of regional pain (see table 2.2).

2.1.2 The clinical relevance of canal measurements

Measurements of the dural-sac

The extracted results of studies relating to quantitative measurement of the dural-sac are presented in Appendix B, tables B.2, B.3 and B.4.

The DS-CSA was the most commonly used quantitative measurement of the dural-sac. The DS-CSA was smaller on average in symptomatic stenosis patients compared to controls, both at the site of maximal stenosis (Y. U. Kim et al. 2015; Schizas et al. 2010), and on average at each lumbar level (Hamanishi et al. 1994).

Cross-sectional type studies largely failed to demonstrate a relationship between DS-CSA and symptom severity. There was some evidence for a relationship with walking capacity — two papers found a positive correlation between a patients smallest DS-CSA and walking capacity (Barz et al. 2008; Kanno et al. 2012), but several other studies did not reproduce these findings (Moon et al. 2005; Kuittinen et al. 2014c; Sigmundsson et al. 2011; Zeifang et al. 2008). A single study demonstrated a positive correlation with the Japanese Orthopaedic Association score (Kanno et al. 2012), but despite repeated investigation, no significant correlation between DS-CSA and pain scores, the Oswestry disability index, or other measures of disability or quality of life at baseline was found. It was also unclear if there was any relationship between the DS-CSA and duration of

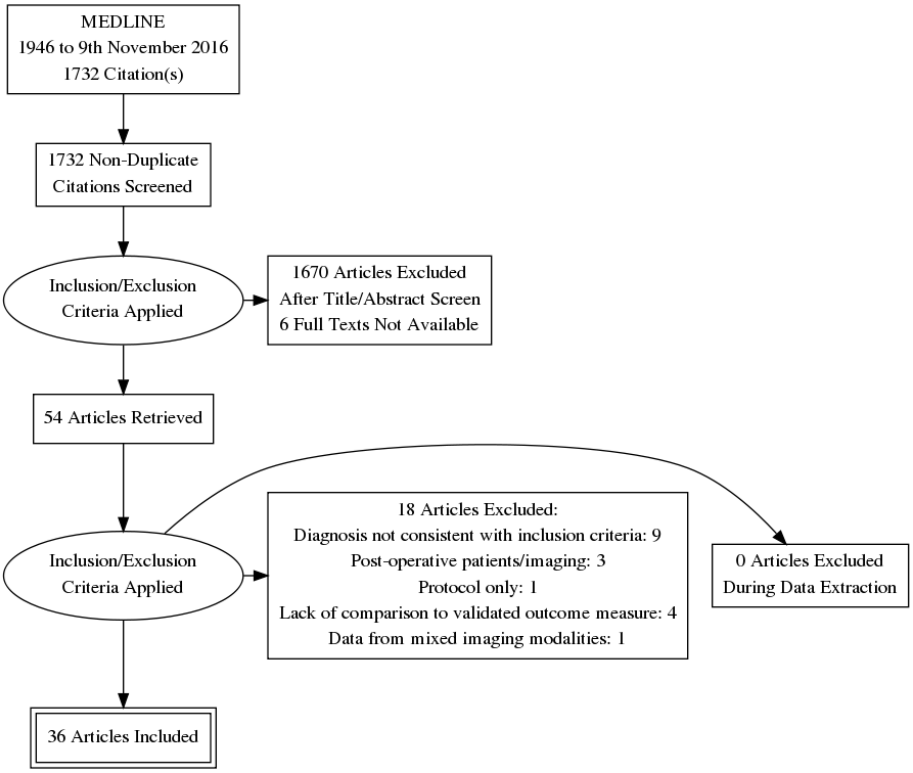


Figure 2.1: PRISMA diagram of scoping search strategy results

Table 2.1: Stratification of included papers by participant source

Participant Source	Frequency	% of Total Papers
Participants undergoing surgical decompression	12	33
Participants attending a spinal clinic	8	22
Participants undergoing epidural steroid injection/adhesionolysis	4	11
Participants recruited from a cohort undergoing MR lumbar spine	3	8
Population cohort	1	3
Undefined	8	22

symptoms, as would be expected if anatomical stenosis is progressive with age, with one paper finding a negative correlation (Kanno et al. 2012), and one finding no relationship (Sirvanci et al. 2008).

Schizas et al. (2010) found a small DS-CSA was predictive of failure of conservative treatment, but did not find an associated correlation with the degree of improvement post surgery, the latter point also confirmed by Kuittinen et al. (2014c) and in post-epidural adhesiolysis by Park C. H. Park et al. (2011). In contrast, Sigmundsson et al. (2011) found some evidence for patients with smaller DS-CSA showing improved outcomes in back and leg pain post operation.

No relation between the dural-sac AP diameter with clinical findings or prognosis was identified (table B.2 and B.4).

No clear relationship between any of the qualitative grading systems measuring dural sac compression and baseline symptoms was demonstrated — a single paper showed more severe grades were associated with higher maximum walking distances and lower pain (Kuittinen et al. 2014c), reversing the limited trend seen in quantitative studies. In contrast, in prognostic studies, there was a trend for more severe stenosis to be related to better post-operative outcome (Kuittinen et al. 2014b; Moojen et al. 2018).

Two qualitative indirect measures of dural-sac compression were assessed in the included papers: The sedimentation sign (failure of the nerve roots to sediment to the dorsal aspect of the dural-sac above the level of the stenosis, table B.6) was found to be relatively sensitive and specific for the presence of symptomatic stenosis (Barz et al. 2008; Macedo et al. 2013b). However these papers separated symptomatic and control patients based on both clinical

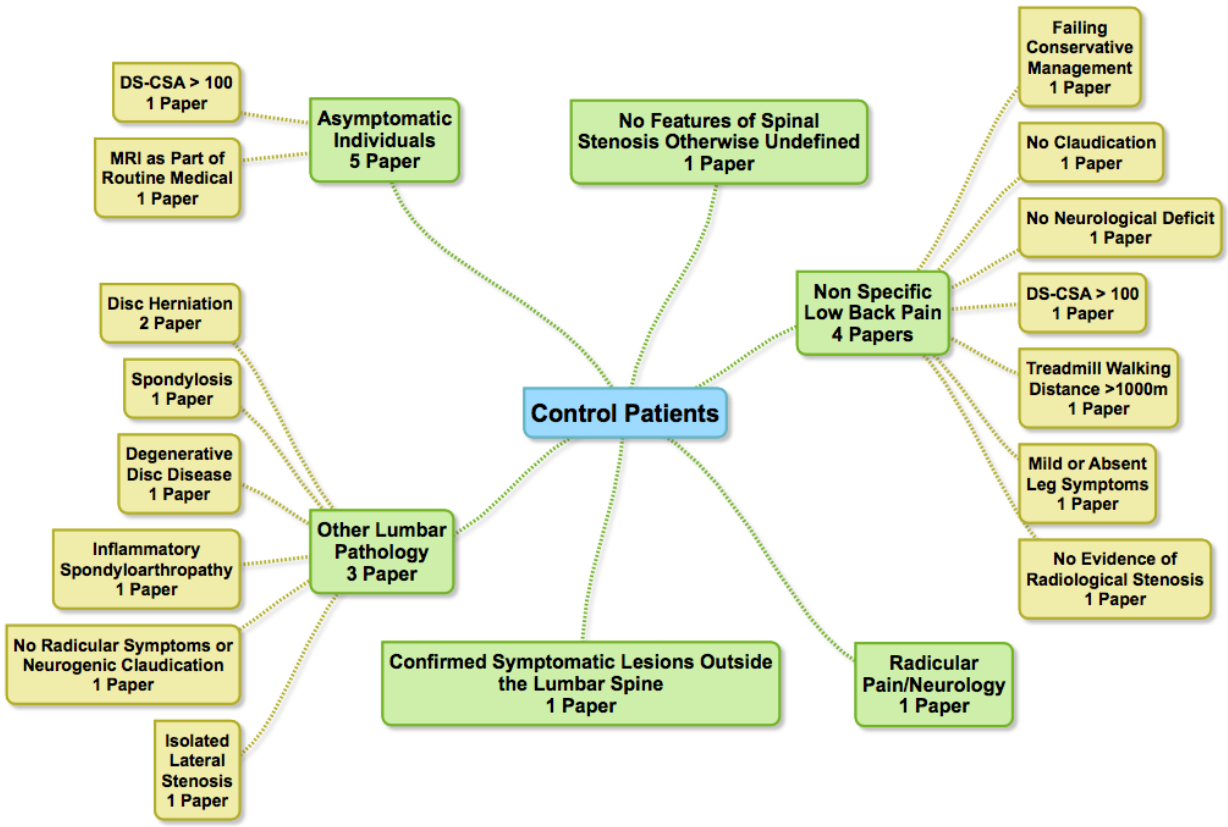


Figure 2.2: Map of the definition of the control group within the included papers inclusion criteria

symptoms and radiological differences, significantly limiting the interpretation of these findings. The presence of the sedimentation sign was linked to poor outcomes in those managed non-operatively (Barz et al. 2014), with increased surgical treatment effect (Barz et al. 2014; Moses et al. 2015). The redundant nerve root sign (the presence of elongation of the nerve roots above the level of the stenosis), showed similar clinical associations (Min et al. 2008; Moojen et al. 2018) (table B.7).

Measurement of the osseous and ligamentous boundaries of the spinal canal

The results of included papers assessing the bony and ligamentous boundaries of the spinal canal using quantitative measurements and qualitative grading systems are detailed in table B.8 and B.9 respectively.

Quantitative measurements generally showed no relation with clinical status, with no convincing correlation between either the canal cross-sectional area, AP diameter, lateral recess height or foraminal cross-sectional area and clinical symptoms at baseline. No case-control or prognostic studies were found for the quantitative measures.

While Ishimoto et al. (2013) found a relationship between qualitatively measured compromise of the central canal and prevalence of symptomatic stenosis in a population cohort, other studies using the same grading system failed to demonstrate a relationship with pain symptoms in stenosis patients (Burgstaller et al. 2016). The absence of epidural fat, indicative of more severe stenosis and a favoured criterion for practising radiologists (Mamisch et al. 2012), was found not to predict pain or disability but did predict a higher rate of satisfactory response to surgery (Burgstaller et al. 2016).

Qualitative grading schemes of the lateral recess and neural exit foramina stenosis focused on whether the transiting or exiting nerve root appeared compressed (table B.10). No significant association with clinical symptoms at baseline or prognosis was identified. No quantitative measurement of lateral recess or neural exit foramen stenosis was investigated by the included papers.

The clinical relevance of multi-level stenosis

Due to the possible pathophysiological importance of multi-level stenosis in the generation of neurogenic claudication, it is surprising only a single relevant case-control type paper was identified. Hamanishi et al. (1994) demonstrated that while 89% of patients with neurogenic claudication appeared to have multi-level central stenosis, this was seen in only 10% of controls.

Paradoxically, two papers found multi-level stenosis patients had less severe symptoms at baseline compared to patients with single level stenosis (table B.11). Sigmundsson et al. (2011) found the number of stenotic levels showed a negative correl-

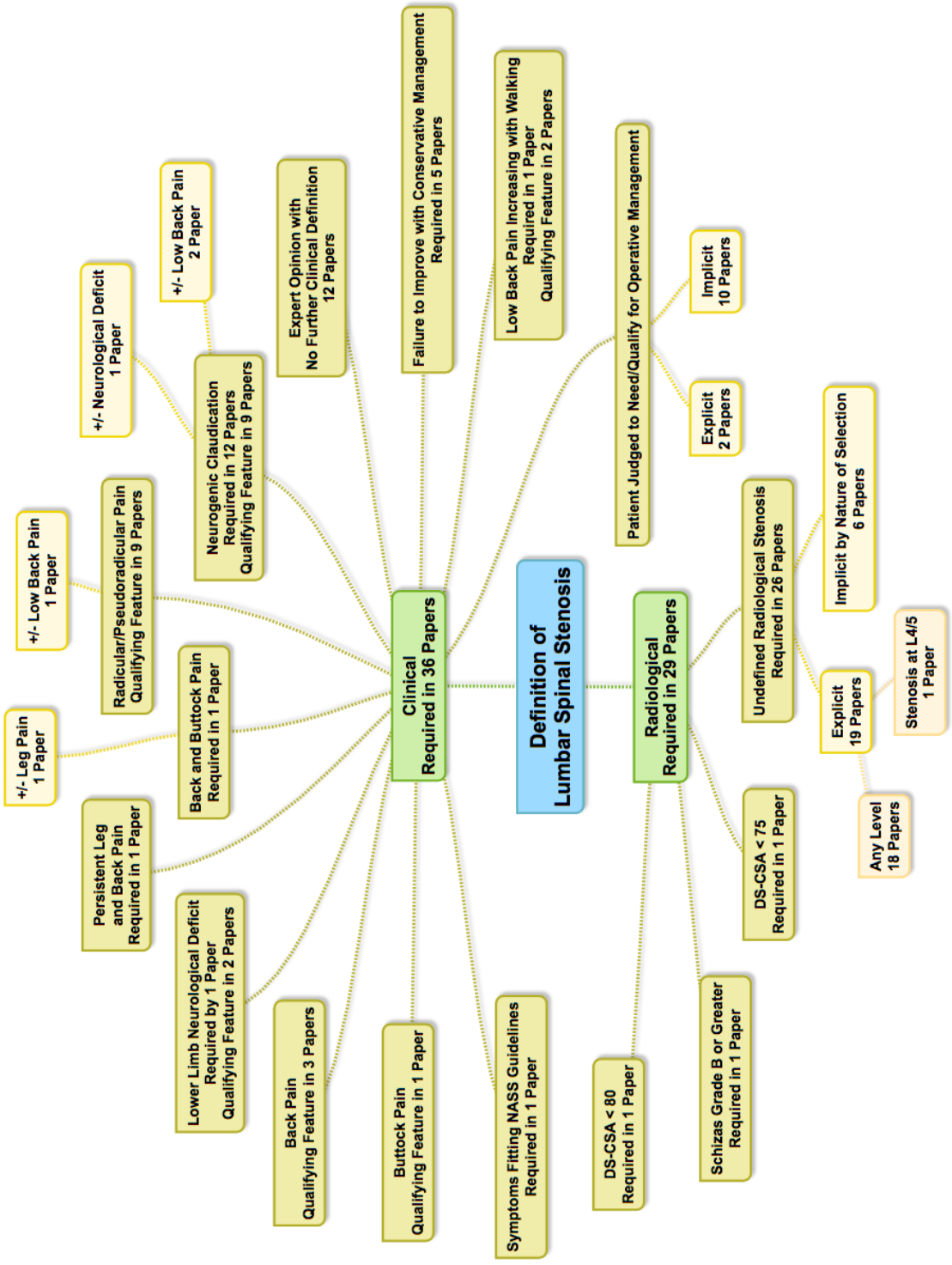


Figure 2.3: Map of the definition of lumbar spinal stenosis within the included papers inclusion criteria

Table 2.2: The use of validated clinical scoring systems for measurement of pain and disability within the papers included in the scoping review. Freq. — Frequency, % — Percentage of total citations.

Clinical Scoring System	Freq.	%	Citation
Oswestry Disability Index	15	42	Fairbank et al. 2000
Visual Analogue Scale for Leg Pain	12	33	Zanoli et al. 2001
Visual Analogue Scale for Back Pain	11	31	Zanoli et al. 2001
Short Form Health Survey (SF-36)	5	14	Ware et al. 1992
Numerical Rating Scale for Back Pain	4	11	Downie et al. 1978
Beck Depression Inventory	3	8	Beck 1961
Lumbar Spinal Stenosis Surgery Satisfaction Score	3	8	Kuittinen et al. 2012
Japanese Orthopaedic Association Scale	3	8	Azimi et al. 2012
Euro-Qol - 5 Dimensions	2	6	EuroQol Group 1990
Numerical Rating Scale for Leg Pain	2	6	Downie et al. 1978
Roland-Morris Score	2	6	Roland et al. 1983
Stenosis Bothersomeness Index	2	6	Patrick et al. 1995
Visual Analogue Scale for Whole Bodily Pain	2	6	Zanoli et al. 2001
Centre for Epidemiologic Studies Depression Scale	1	3	Carleton et al. 2013
Hannover Back Pain Activity Score	1	3	Kohlmann et al. 1994
McGill Pain Questionnaire	1	3	Melzack 1975
Qúebec Back Pain Disability Scale	1	3	Kopec et al. 1995
Pain Disability Index	1	3	Tait et al. 1990
Pain Sensitivity Questionnaire	1	3	Ruscheweyh et al. 2009
Panic Disorder Severity Scale	1	3	Houck et al. 2002
Pittsburgh Sleep Quality Index	1	3	Mollayeva et al. 2016
Swiss Spinal Stenosis Questionnaire	1	3	Stucki et al. 1996

ation with VAS leg pain and that patients with multi-level stenosis had higher general health levels. Kuittinen et al. (2014c) found longer maximum walking distances in multi-level stenosis. These authors hypothesised that multi-level facet joint arthrosis may promote stability, reducing spinal instability related symptoms. Other studies have not reproduced these findings (Sirvanci et al. 2008; Kapural et al. 2007; Zeifang et al. 2008). The prognostic value of the presence of multi-level stenosis was unclear (table B.12), with two studies showing improved surgical treatment effects in single level stenosis (D. K. Park et al. 2010; Kuittinen et al. 2014b) and one showing the opposite (Sigmundsson et al. 2012).

2.2 Systematic review rationale and methodology

Aims and rationale

At the conclusion of the scoping review, two review questions relevant to the research questions posed in § 1.7.2 were selected for full systematic reviews.

Review 1: primary review question:

> In patients diagnosed with degenerative lumbar spinal stenosis, what is the relationship between measurements of the dural-sac cross-sectional area on MRIs of the lumbar spine, and the patient's pain, disability and quality of life?

The DS-CSA was selected as it was by far the most common measurement of anatomical LSS severity in the reviewed literature, measured a structure directly linked to the risk of nerve root compression, and was investigated by papers with a relatively homogenous methodology but with unclear aggregate results.

A review of the relationship between DS-CSA and the efficacy of surgical intervention, while perhaps clinically more interesting, was not attempted due to the small number of papers investigating this and their significant methodological heterogeneity both in terms of intervention and statistical approach.

Review 2: primary review question:

> What quantitative or qualitative findings on MRI examinations of the lumbar spine are associated with the presence of neurogenic claudication?

The small number of case-control studies identified by the scoping review generally found smaller canal sizes and indirect signs of neural compression in patients with symptomatic LSS, but used heterogenous inclusion criteria — not all adhered to the NASS definition or explicitly required patients to have neurogenic claudication. Hence, a systematic review of case-control papers using a case population specifically with neurogenic claudication, and hence closer to the clinical side of the NASS definition, was conducted.

Prospective registration

The protocol for the two subsequently described systematic reviews were prospectively registered on the PROSPERO international register of systematic reviews under references: CRD42017064865 (Gagen et al. 2017a) and CRD42017064873 (Gagen et al. 2017b).

Search strategy

A combined search strategy was used for both systematic reviews, designed to return citations referring to MRI imaging in patients with spinal stenosis and/or neurogenic claudication. The search was performed on MEDLINE, EMBASE, the Cochrane Library and Web of Science databases on the 23rd August 2017 with no specified start date. The full search strategy is included in Appendix B.2.

References were managed using the website http://www.covidence.org (Veritas Health Innovation 2017) and duplicates were removed. Citations were screened first by title and abstract, with separation of citations relevant to the different reviews at the abstract screening phase. Identified potentially relevant citations then had their full texts by two authors. Disputes regarding inclusion were settled by a third author. No language limitations were imposed, two non-English papers were translated during full text screening. The bibliography of included citations were assessed for any citations missed by the search strategy.

No missing full texts were identified. The reasons for the exclusion of any papers at the full text review were recorded.

Eligibility criteria

Citations generated by the search strategy were screened according to the following eligibility criteria.

Review 1 inclusion criteria:

- The studied population must consist of skeletally mature adults with a primary diagnosis of symptomatic degenerative lumbar spinal stenosis.
- Included papers must report the dural sac cross-sectional area, as measured on cross-sectional imaging, and relate this to a validated and quantitative metric of pain, disability or quality of life.
- Included studies will be primarily cross-sectional in design or have a cross-sectional component and be published within a peer reviewed journal or as a conference proceeding.

Review 2 inclusion criteria:

- The studied population within included citations must:
  - Involve only skeletally mature adults.
  - Include a case population who on clinical assessment have neurogenic claudication defined as radiating buttock or leg pain that is provoked by upright exercise.
  - Include at least one control population who do not have neurogenic claudication symptoms.
- Report at least one quantitative or qualitative finding on MRI examination of the lumbar spine and compare the presence of this finding between the neurogenic claudication and control groups.

It is important to note that for review two, studies that use a "spinal stenosis" case population, where neurogenic claudication is not stated to be present in all case participants, would not meet the above inclusion criteria.

Exclusion criteria (both reviews):

- Studies specifically including patients with the conditions below, where these patients are not analysed separately:
  - Acute spinal conditions (e.g. disk herniation, trauma, infection)
  - Spinal malignancy
  - Prior lumbar spinal surgery
  - Congenital syndromes

Data extraction

Papers identified for inclusion underwent data extraction by two authors using a standardised template. The data extracted included a description of the study population (including age range, ethnicity, the inclusion and exclusion criteria and recruitment method) and a description of the study methodology (including MRI positioning, MRI sequences employed, the method used for measurement of the dural-sac cross-sectional area and the timing of the data collection). The specific outcome variables, their means of collection, the statistical methods employed and the results of each analysis presented for each paper was also recorded.

Quality assessment was performed for each included paper, guided by the National Institute of Health (NIH) quality assessment tool for observational cohort and cross-sectional studies for review 1 (questions 6, 7, 10 and 13 were omitted as they are only relevant to longitudinal studies) and the NIH quality assessment tool for case-control studies for review 2 as appropriate to the expected included study type. Quality assessment was completed by two authors for each paper and used to inform an overall assessment of the risk of bias as being either good, fair or poor. No specific quality threshold was used to exclude papers from any subsequent analysis, and a discussion of any identified sources of bias and study quality is included in the review.

After completion of data extraction, discrepancies were resolved by discussion among collaborators.

Data synthesis and meta-analysis

Primary data synthesis is tabular with narrative summary and discussion of the results, broken down by outcome variable, for both reviews.

For review 1, the majority of papers performed a correlation analysis between the minimal DS-CSA per patient and the various outcome variables. Where appropriate, a meta-analysis of correlation coefficients has been performed using a similar approach to the "bare bones" method proposed by Schmidt et al. (2015). For outcome variables where no meta-analysis has been performed, the reasoning behind the decision not to proceed has been presented.

Two papers included in review 1 presented a comparison between DS-CSA and ODI grouped as categorical variables — the data from these two papers was extracted, pooled, and a chi-squared test, using a null hypothesis that the DS-CSA category and ODI category were unrelated, was performed.

One included paper within review 1 (Weiner et al. 2007) reported both DS-CSA and an outcome variable for each participant but did not perform any statistical analysis between these variables. This data was extracted and a Pearson's correlation coefficient calculated.

Within the text descriptions of correlation coefficients adhere to Cohen's conventions for effect sizes (small: 0.1 – 0.3, medium: 0.3 – 0.5, large: > 0.5). Stated p values and other statistics are taken from the original papers where possible, or calculated where absent. Missing p values for correlation analyses were calculated using a Fisher's Z transformation, and relate to a null hypothesis of no true correlation between DS-CSA and the outcome variable.

For review 2, no meta-analysis of study results relating to diagnostic test accuracy was performed due to both high methodological heterogeneity and problems with bias within study results, as discussed further in § 2.4.

2.3 Review 1 results:

Is there a relationship between the dural-sac cross-sectional area and patient pain, disability and quality of life?

2.3.1 Search results

The results of the search strategy are detailed in figure 2.4. In total 5155 non-duplicate citations underwent title and abstract screening.

A total of 75 full texts were reviewed and 57 were excluded. The reasons for exclusion at the full text review stage are detailed in the PRISMA diagram. A single paper was excluded during data extraction as it did not state the direction of identified correlations between the DS-CSA and the outcome variables, meaning its results could not be meaningfully incorporated into the review.

2.3.2 Quality assessment

The results of the quality assessment are shown in figure 2.5. In total 12 of the 17 papers were judged to be of poor quality with 3 being judged as fair and 1 judged of good quality.

There were several domains in which papers consistently scored poorly. Only one paper made an attempt to statistically control for potential confounding variables and few made any assessment of their statistical power. Many studies used radiological criteria as part of their inclusion criteria, either implicitly or explicitly, but did not state the definition of stenosis they used to include patients. There was also generally poor recording of whether MRI and clinical assessors were appropriately blinded and many papers did not assess the reliability of their measurement of the DS-CSA.

2.3.3 Relationship with disease severity

The majority of papers assessed the relationship between the minimal DS-CSA per patient and one or more outcome variables. A summary of the findings related to minimal DS-CSA are shown in figure 2.6 as a harvest plot (Ogilvie et al. 2008). The results of pooling of correlation coefficients are shown in table 2.3.

Walking capacity

In total 8 papers attempted to find a relationship between the DS-CSA and walking capacity. A textural summary of the results of these papers is shown in table 2.4.

Prasad et al. (2016) and Barz et al. (2008) found moderate positive correlations between minimal DS-CSA per patient and treadmill based maximum walking distance (MWD). In contrast, Zeifang et al. (2008) and Moon et al. (2005) found no correlation. Zeifang et al. (2008) did find a small negative correlation between DS-CSA at L1–2 and treadmill MWD but this was not replicated at any other lumbar level.

Three papers assessed patient reported MWD. Y. U. Kim et al. (2015) and Kanno et al. (2012) found small positive correlations with minimal DS-CSA per patient. Interestingly, Kanno et al. (2012) found a stronger correlation on axial loaded MRI compared to standard supine MRI. Sigmundsson et al. (2011) found no significant correlation.

Lau et al. (2017) mixed patient self-report and GPS recorded maximum walking distance and found a moderate positive correlation with the minimal DS-CSA. Similar to the findings of Kanno et al. (2012) the strength of this correlation increased on standing MRI.

No clear relationship between study quality or sample size and the significance or size of reported correlations was identified.

Pooling of correlation coefficients was performed for papers relating the minimal DS-CSA to MWD.

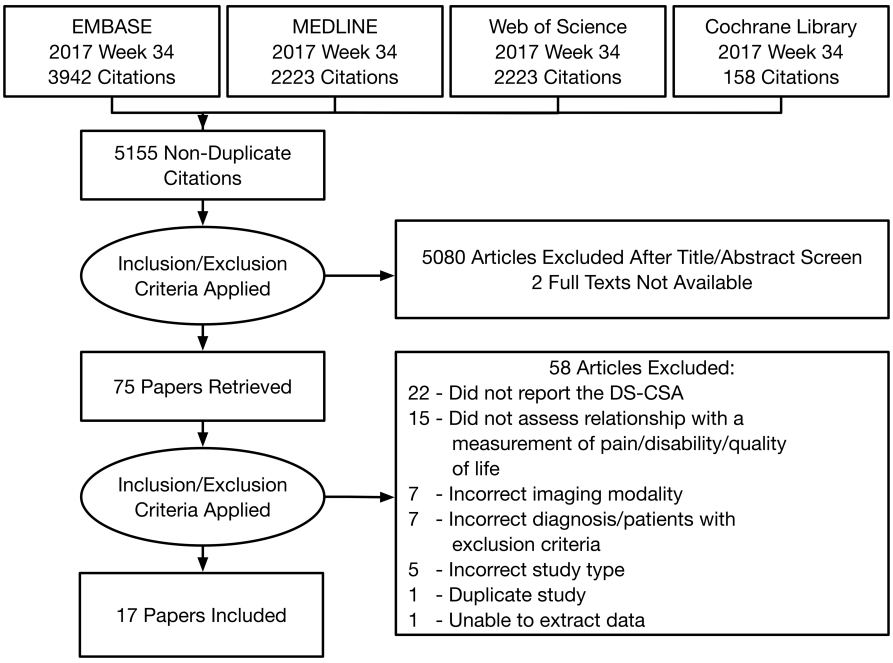


Figure 2.4: PRISMA Diagram of Search Strategy Results

Table 2.3: Presentation of the results of pooling correlation coefficients between minimal DS-CSA per patient and outcome variables relating to disease severity.

Outcome	Studies	Total N	Pooled Cor.	95% CI			p	I^2
Max. Walking Distance (Overall)	8	551	0.14	−0.02	–	0.30	0.08	69
Max. Walking Distance (Treadmill)	4	171	0.05	−0.27	–	0.38	0.74	75
Max. Walking Distance (Patient Report)	3	310	0.14	0.02	–	0.26	0.03	17
Walking Time to Symptom Onset	2	83	0.06	−0.19	–	0.31	0.65	22
SF-36 (Bodily Pain Domain)	2	182	0.05	−0.09	–	0.20	0.47	0
SF-36 (General Health Domain)	2	182	−0.06	−0.21	–	0.08	0.41	0
SF-36 (Mental Health Domain)	2	118	0.07	−0.12	–	0.25	0.48	0
SF-36 (Physical Function Domain)	3	230	0.04	−0.09	–	0.17	0.53	0
SF-36 (Role Physical Domain)	2	182	0.12	−0.02	–	0.26	0.10	0
Overall Pain	3	248	0.06	−0.06	–	0.19	0.33	0
JOA Score	2	136	0.09	−0.27	–	0.44	0.63	74

The overall random effects model (figure 2.7) did not reach significance with a pooled correlation coefficient of 0.14 (95% CI: -0.02–0.30, p = 0.08). Significant statistical heterogeneity was identified (I^2 69.27%). In order to investigate this, an additional analysis was performed producing separate models for papers measuring maximum walking distance via treadmill examination and via subjective patient report. The model for subjective walking distance just reached significance with a pooled correlation coefficient of 0.14 (95% CI: 0.02–0.26, p = 0.025) while the model for treadmill walking distance remained non-significant with significant heterogeneity (I^2 72.45%). A funnel plot for this data was symmetric (figure 2.8).

The studies using patient self-report were generally larger and of higher quality (mean n = 103, two of three ranked fair quality) compared to the studies using treadmill assessment (mean n = 42, one of four ranked fair quality) potentially explaining the differences in heterogeneity between the subgroups. All papers used secondary care populations, most being patients managed surgically.

Two papers assessed the relationship between DS-CSA and walking distance/time to symptom onset but neither found a significant relationship (Prasad et al. 2016; Moon et al. 2005). Sigmundsson et al. (2011) found no significant difference in patient reported walking distance between single and multi-level stenosis using a DS-CSA of < 70 mm^2 as a threshold for diagnosing stenosis

Oswestry disability index (ODI)

The ODI is a self completed questionnaire designed for assessment of function and disability in individuals with low back pain (Fairbank et al. 2000). The index ranges between 0 and 100 with higher numbers representing greater disability.

In total 10 studies assessed the relationship between the ODI and DS-CSA. A summary of the results of these studies is shown in table B.21.

Of the 9 studies performing correlational analyses between minimal DS-CSA and ODI, no study found a significant correlation. In addition, Lau et al. (2017) found no significant correlation with minimal DS-CSA on standing MRI. Six papers did not report

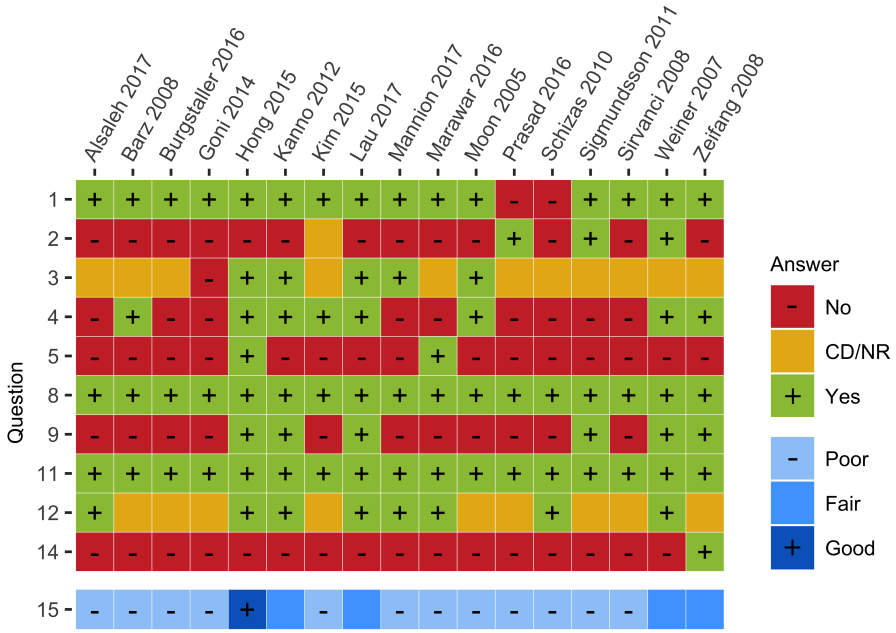


Figure 2.5: Quality Assessment as per NIH quality assessment tool for observational cohort and cross-sectional studies – Questions 1 to 14 relate to specific aspects of paper quality and are answered as yes, no or CD/NR (cannot determine/not recorded). Question 15 relates to overall paper quality and is judged as good, fair or poor. "+" symbols are used to highlight answers of yes/good and "-" symbols are used to highlight answers of no/poor.

the correlation co-efficient and one did not include information on sample size. Given this, and other study heterogeneity, no meta-analysis was therefore attempted for this data.

Two papers categorised patients by ODI into minimal disability (0% to 20%), moderate disability (21% to 40%), severe disability (41% to 60%), crippled (61% to 80%) and bedridden (81% to 100%) groups and compared the distribution of DS-CSA between them. Goni et al. (2014) found significantly more patients with a minimal DS-CSA < 70 mm^2 in higher level disability groups compared to patients with minimal DS-CSA > 70 mm^2 ($p = 0.026$). In contrast, when patients were categorised by DS-CSA as having severe stenosis (<76 mm^2), moderate stenosis (76 to 100 mm^2) or as normal (>100 mm^2) no relationship between DS-CSA category and ODI category was identified. The latter finding was reproduced by Sirvanci et al. (2008) and pooling the results from these two papers did not alter the significance of the results (Chi-squared = 8.9, $p = 0.37$).

Hong et al. (2015) found a small positive correlation between the number of levels with DS-CSA $<$ 100 mm^2 and the ODI ($p < 0.045$). Similarly dividing patients into single and multilevel stenosis groups showed that the ODI score was significantly smaller in the single level stenosis group ($p = 0.040$).

Short form 36 (SF-36)

The Short-Form 36 questionnaire (SF-36) is a questionnaire designed for group comparisons, not specific to any one disease (Ware et al. 1992). It measures both physical and mental health components across 8 domains. The index ranges from 0 to 100 with larger numbers representing higher quality of life.

A summary of papers relating to SF-36 is shown in table B.22. Five studies assessed the relationship between minimal DS-CSA per patient with selected domains of the SF-36, but none identified significant correlations. Pooling of these results separately for each domain did not result in any of these relationships becoming significant. Alsaleh et al. (2017) found no significant correlation between the minimal DS-CSA per patient and the total SF-36 score.

Sigmundsson et al. (2011) found that patients with multi-level stenosis (defined as more than one level with a DS-CSA < 70 mm^2) had significantly higher SF-36 general health scores compared to single level stenosis groups ($p = 0.04$). They did not find significant differences in bodily pain, physical function or role physical domain scores. In contrast, Hong et al. (2015) comparing the same SF-36 domains between single and multilevel stenosis groups using a DS-CSA < 100 mm^2 threshold found no significant differences.

Lau et al. (2017) did not find any significant relationship between minimal DS-CSA and physical health or mental health domains on standing MRI.

Pain severity

VAS allow assessment of patient reported symptom severity by asking the patient to mark a line between 0 and 100 mm. Marks closer to the 100 mm mark represent higher symptom severity, typically measured over the previous week. Similarly, the numeric rating scale (NRS) technique asks participants to rate pain on integer scale from 0 to 10, rather using the continuous line of the VAS method. Both are known to produce comparable results (Downie et al. 1978). Overall, 9 studies assessed the relationship between

Figure 2.6: Harvest plot of the included papers findings with regards to the relationship between the minimal DS-CSA and the various outcome variables. Each papers is represented by a bar, placed within the grid for its outcome variable assessed and the relationship identified. The height of each bar is proportional to the papers sample size which is written above the bar. The bars colour relates to the papers overall quality rating. JOA – Japanese Orthopaedic Association, NCO – neurogenic claudication outcome, SF-36 – short form 36, VAS – visual analogue scale, NRS – numeric rating scale.

Table 2.4: Table of study results related to assessment of walking ability. WDAM – Walking distance assessment method. PSR — Patient self-report. TWA — treadmill walking assessment. GPS – mixed GPS and patient self-report.

WDAM	Reference	Study Population	Results
PSR	Sigmundsson 2011	105 Surgically Managed Patients	No significant correlation of minimal DS-CSA or number of levels with DS-CSA <70 mm2 and maximal walking distance. No significant difference in walking distance between patients with two level (n = 54) and single level stenosis (stenosis defined as DS-CSA <70 mm2).
	Kanno 2012	88 Referrals for Surgical Management	Small positive correlation between minimal DS-CSA and maximal walking distance (r = 0.23 p <0.05). After axial loading the correlation increased to 0.46 (p <0.001).
	Kim 2015	117 Secondary Care Outpatients	Small positive correlation between minimal DS-CSA and maximal walking distance (r = 0.201 p = 0.049).
TWA	Moon 2005	35 Surgically Managed Patients	No significant correlation between minimal DS-CSA and maximum walking distance or distance to symptom onset.
	Barz 2008	25 Surgically Managed Patients	Large positive correlation between minimal DS-CSA and maximum walking distance (r = 0.53 p = 0.003).
	Zeifang 2008	63 Secondary Care Outpatients	Small negative correlation between DS-CSA at L1–2 and maximal walking distance (tau = -0.118 p = 0.032). No significant correlation at any other lumbar level or with minimal or mean DS-CSA per patient.
	Prasad 2016	48 Surgically Managed Patients	Moderate positive correlation between minimal DS-CSA and maximum walking distance (r = 0.35 p = 0.01) and maximum walking time (r = 0.45 p <0.01). No significant correlation between minimal DS-CSA and distance or time to symptom onset.
MGPS	Lau 2017	70 Secondary Care Outpatients	Moderate positive correlation between minimal DS-CSA and walking distance (r 0.39 p <0.001). Correlation increased to 0.55 (p <0.001) on standing MRI.

DS-CSA and VAS or NRS assessment of symptom severity (table B.23).

Three studies assessed overall pain severity, two using VAS, one using NRS scores. All three studies found no significant correlation between the minimal DS-CSA and overall pain rating. The pooled correlation co-efficient for these studies was not significant. Burgstaller et al. (2016) also found no significant correlation between DS-CSA on a per segment basis and overall pain. Hong et al. (2015), using a DS-CSA threshold of 100 mm^2 to define stenosis, found no correlations between the number of stenotic levels and pain levels and no difference in average overall pain between single and multi-level stenosis groups.

Three studies assessed low back pain symptom severity, all three using VAS scores, and none finding a significant correlation with minimal DS-CSA per patient. (Sigmundsson et al. 2011) found no significant correlation between VAS back pain scores and the number of levels with DS-CSA < 70 mm^2 per patient. No significant difference between single and multi-level stenosis groups was identified using DS-CSA $<$ 70 mm^2 as a threshold.

Six studies assessed leg symptom severity, five assessing leg pain and one assessing leg numbness using VAS scores. On standard supine MRI no paper found a significant correlation between minimal DS-CSA and VAS leg pain. On standing MRI, Lau et al. (2017) found a small negative correlation with VAS leg pain ($p < 0.05$). In contrast, Kanno et al. (2012) did not find a similar relationship using axially loaded MRI. Sigmundsson et al. (2011) found a small negative correlation between the number of levels with DS-CSA < 70 and VAS leg pain ($p = 0.03$) but when dividing patients into single and multi-level stenosis groups, using the DS-CSA < 70 mm^2 threshold, found no significant difference between groups.

Kanno et al. (2012) assessed VAS scores for leg numbness and found no significant correlation with minimal DS-CSA on supine MRI. On axially loaded MRI, however, a small negative correlation was identified ($p < 0.01$).

Due to incomplete data no attempt to pool correlation coefficients relating to low back pain or leg symptom severity was made.

Japanese orthopaedic association score

The Japanese Orthopaedic Association (JOA) score is a 7 item metric based on a mixture of patient report and examination findings (Fujiwara et al. 2003). Lower scores represent poorer outcomes. Two papers in total assessed JOA score, summarised in table B.24.

Kanno et al. (2012) found a small positive correlation between the minimal DS-CSA and the JOA score ($p < 0.05$) on supine MRI. This increased to a moderate correlation on axial loading ($p < 0.001$). Prasad et al. (2016) did not find a significant correlation on supine MRI, albeit with a sample size almost half that

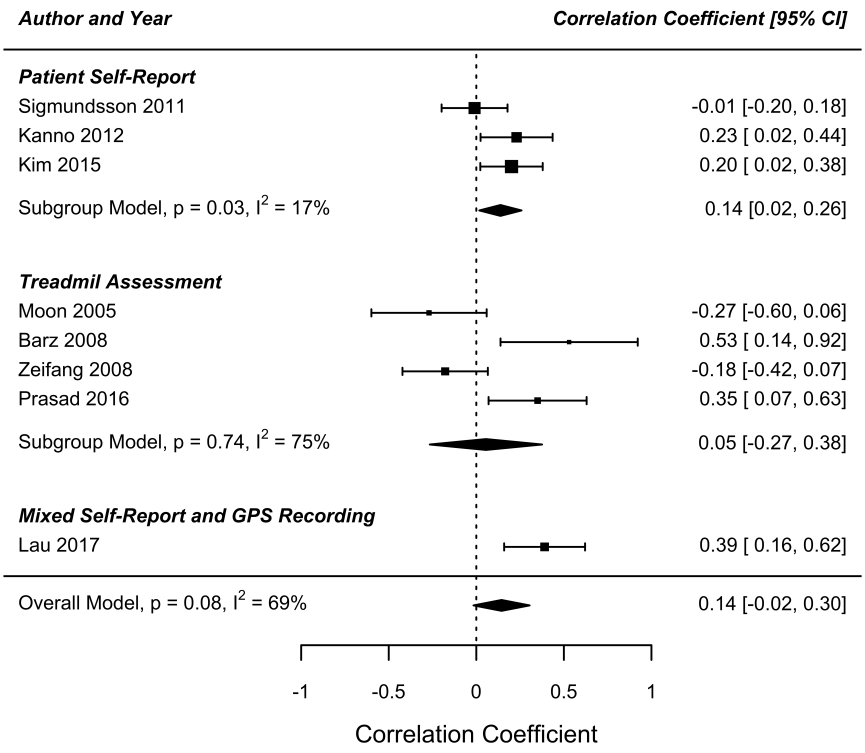


Figure 2.7: Forrest plot of correlational analyses relating minimal DS-CSA to maximum patient walking distance. The identified correlation co-efficient is presented for each study with the upper and lower confidence intervals.

of Kanno et al. (2012). The pooled correlation coefficient of 0.09 (CI: -0.27 – 0.44) was not significant (p = 0.63).

Zurich claudication questionnaire

The Zurich claudication questionnaire (ZCQ) is a disease specific self report outcome instrument commonly used in trials of treatment of lumbar spinal stenosis(Stucki et al. 1996). The scale refers to symptoms over the past month and is expressed as a percentage with higher values representing worsening disability.

Two studies assessed the ZCQ, summarised in table B.24. Burgstaller et al. (2016) found no significant correlation between minimal DS-CSA and the symptom severity domain of the ZCQ. Marawar et al. (2016) found a positive correlation between the DS-CSA at L2–3 and ZCQ scores (p value not stated) that was not present at any other level.

Other outcome measures

Hong et al. (2015) found no significant correlation between minimal DS-CSA or the number of levels with DS-CSA < 100 mm^2 and the Beck depression, Beck anxiety or Pittsburgh sleep quality indices. No difference in these outcome variables was identified between single and multi-level stenosis groups. Mannion et al. (2017) found no significant relationship

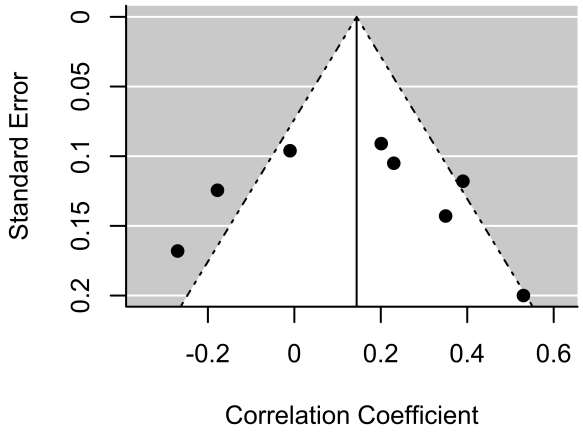


Figure 2.8: Funnel plot of correlation coefficients between minimal DS-CSA and maximum walking distance.

between DS-CSA with the core outcomes measure index (a low back pain assessment tool (Azimi et al. 2012)). Sigmundsson et al. (2011) found no significant relationship between EQ-5D scores and minimal DS-CSA or the number of levels with DS-CSA < 70 mm^2. No significant difference in average EQ-5D scores was found between single and multi-level stenosis groups. Weiner et al. (2007) found no significant correlation between the minimal DS-CSA per patient and the neurogenic claudication outcome score.

2.4 Review 2 results:

What MRI findings are associated with neurogenic claudication?

2.4.1 Search results

The results of the search strategy are detailed in figure 2.9. 5155 non-duplicate citations underwent title and abstract screening.

A total of 57 full texts were reviewed and 43 were excluded. Two citations were excluded as their full texts were not available for review. No additional citations were identified on review of the included papers bibliographies.

2.4.2 Quality assessment

The results of the quality assessment are shown in figure 2.10.

A number of consistent methodological problems were identified in the included papers. All but one study was designed primarily to investigate symptomatic lumbar spinal stenosis, rather than specifically patients with neurogenic claudication, and hence included radiological criteria within their case participant definition. This was sometimes explicit (though the definition of stenosis was often not provided), but also occurred through use of a case population drawn from surgically managed patients. Eleven studies did not blind MRI assessors to the group status of the participants. No included study provided a sample size justification and only a single paper attempted to statistically control for potential confounding variables.

2.4.3 MRI features

Dural-sac cross-sectional area

Four papers reported the DS-CSA in both case and control participant groups (table A.10).

Hamanishi et al. (1994) reported the minimal DS-CSA was smaller at all measured levels in neurogenic claudication patients when compared to a small age matched control group. The majority of neurogenic claudication patients showed stenosis at the L4/5 level with a mean DS-CSA of 47 ± 19 mm^2 compared to 144 ± 35 mm ^{2} in controls. For the 134 participants for which MRI images were available between L2/3 and L4/5, 89.5% of claudication patients had two or more levels with a DS-CSA < 100 mm^2, compared to 22.9% of radicular pain patients and 3.1% of low back pain patients.

The remaining three studies all found smaller DS-CSA measurements in case participants, but included radiological criteria within their case participant definition, making interpretation of their findings difficult.

The sedimentation sign

The sedimentation sign, failure of the nerve roots to sediment into the bottom half of the spinal canal adjacent to a stenotic segment, was investigated by four papers (table B.17). All four investigated case participants included radiological criteria as part of their case definition. They reported sensitivities between 94% and 42% and specificities between 100% and 49%.

The highest sensitivity and specificity was reported by Barz et al. (2010) in 2010 who used strict radiological definitions for both case (DS-CSA < 80 mm^2) and control patients (DS-CSA > 120 mm^2). L. Zhang et al. (2017) reported a sensitivity of 77.1% and specificity of 53.0%, using a stenosis definition of a DS-CSA < 120 mm^2, but found this increased to 95% and 80% respectively when applying the case and control definitions used by Barz et al. A logistic regression analysis found the presence of a positive sedimentation sign was not predicted by the presence of neurogenic claudication (OR 0.88, p = 0.56) but was predicted by the presence of radiological stenosis and the number of stenotic levels ($p < 0.01$). Laudato et al. (2015) found a positive sedimentation sign was associated with surgical management within lumbar spinal stenosis groups.

Dural-sac morphology

Two papers assessed the A to D grading system, proposed by Schizas et al. (2010), classifying the effect of lumbar canal stenosis on dural-sac morphology (table B.18). Both papers included both conservatively and surgically managed patients and used a symptomatic lumbar spinal stenosis population, though the definition of radiological stenosis used was not provided.

Schizas et al. (2010) found grades C-D stenosis had a 61.8% sensitivity and 88.8% specificity for the neurogenic claudication group with Laudato et al. (2015) finding 76% and 92% respectively. Both papers found grades C-D stenosis was predictive of failure of conservative management and associated with surgical treatment.

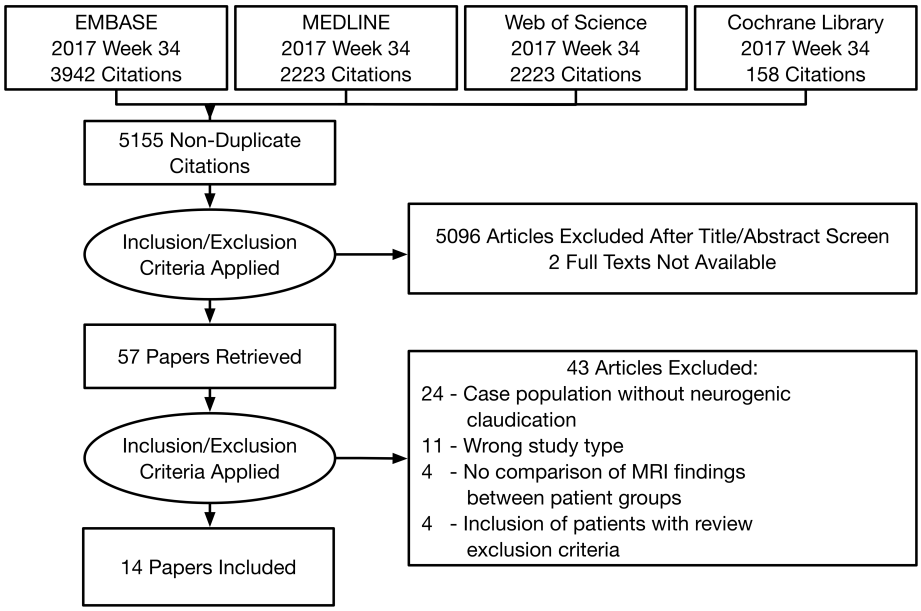


Figure 2.9: PRISMA diagram of search strategy results

Table 2.5: Case-control studies comparing the dural-sac cross-sectional area (DS-CSA)

Citation	Case Population	Control Population	Results
Hamanishi 1994	53 secondary care patients with neurogenic claudication	169 secondary care patients - 51 with non-specific low back pain, 79 with radicular pain and 39 control patients with symptomatic lesions outside the lumbar spine (15 aged matched subgroup).	Mean DS-CSA was smaller at all measured levels in neurogenic claudication patients compared to the age matched control subgroup ($p < 0.001$). 90% of patients with claudication had two or more levels with DS-CSA <100 mm^2, but only 23% of radicular pain patients, 3% of low back pain and 3% of control patients.
Hou 2015	21 patients with neurogenic claudication and a DS-CSA less than two thirds that of a corresponding normal intervertebral space.	20 healthy volunteers	Mean DS-CSA from L3/4 to L5/S1 was significantly smaller in neurogenic claudication patients compared to controls ($p < 0.001$).
Kim 2015	117 pain clinic attenders with neurogenic claudication and radiologically diagnosed stenosis.	91 patients undergoing MRI as part of a routine medical check-up without symptoms of lumbar spinal stenosis.	Minimal DS-CSA per patient was significantly smaller in the neurogenic claudication group compared to the control group ($p < 0.001$).
Zhang 2017	105 orthopedic outpatients with neurogenic claudication and DS-CSA <120 mm^2.	215 orthopedic outpatients with non-specific low back pain.	Minimal DS-CSA was 85.5 mm^2 in LSS patients and 112.2 mm^2 in the low back pain group ($p < 0.01$). 115 of the 215 control patients had radiological stenosis as defined by a DS-CSA <120 mm^2.

Dimensions of the ligament flavum

Two papers assessed measurements of the ligamentum flavum (table B.19). Y. U. Kim et al. (2015) showed individuals with neurogenic claudication and neuroradiologist diagnosed stenosis had greater mean ligamentum flavum area and thickness compared to controls . A second paper by Kim in 2017 showed ligamentum flavum area could separate case and control populations with a sensitivity of 80.1% sensitivity and 76% specificity, diagnostically outperforming ligamentum flavum thickness measurements (Y. U. Kim et al. 2017).

Other MRI measurements

A further 7 MRI measurements were assessed by single papers (see table B.20), all of which included radiological criteria within their case definitions. Manaka et al. (2003) found an increased prevalence of spinal venous plexus defects in claudication patients with higher grades associated with shorter claudication distances. Kobayashi et al. (2015) showed abnormal contrast enhancement of the cauda equina at the site of stenosis in claudication patients. Single papers found a smaller spinal canal area (Y. U. Kim et al. 2015), smaller sacral angle (Ghasemi et al. 2016) and greater multifidus atrophy (Jiang et al. 2017) in claudication patients. Chun et al. (2017) found that after exercise claudication patients showed lower dural peak CSF velocities, which were not present at rest. Hou et al. (2015) found abnormal nerve root tractography in claudication patients.

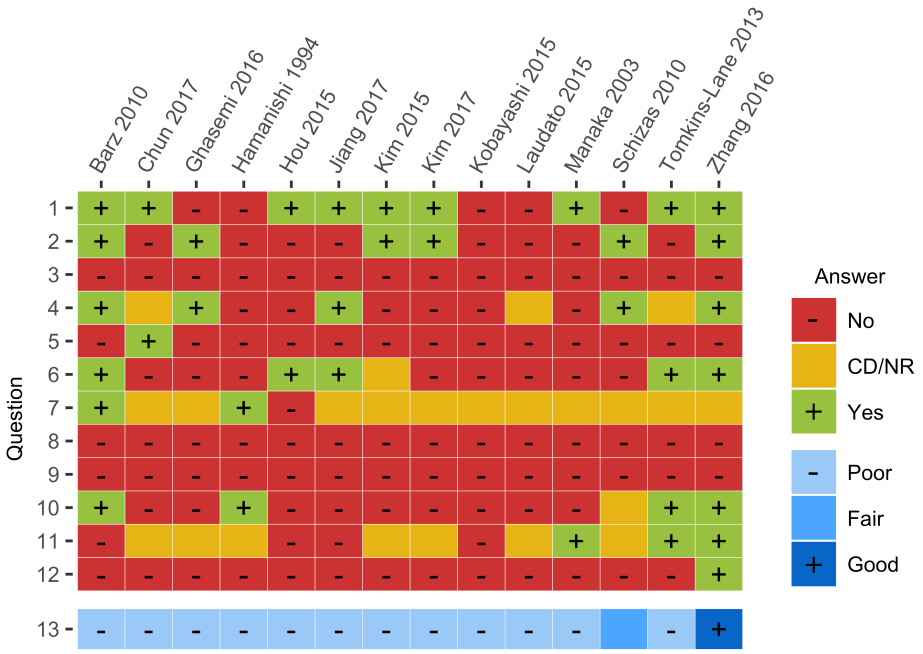


Figure 2.10: Graphical summary of quality assessment for the included papers

2.5 Literature review discussion

2.5.1 Review 1

This review included 17 papers exploring the relationship between the DS-CSA (a measurement of spinal canal narrowing) and multiple measurements of patient pain, disability and quality of life.

The best evidence was found for a relationship with maximum walking distance. A slight majority of papers found a positive relationship between minimal DS-CSA and maximum walking distance but the overall pooled correlation co-efficient was not significant with substantial heterogeneity. The heterogeneity appeared to be primarily caused by a group of studies all using treadmill assessment that had on average smaller sample sizes and lower quality scores than the much more homogeneous studies using patient self-report. A subgroup model pooling the latter group of papers found a small but significant positive correlation, similar in size to that found by Lau et al. (2017) who mixed self-report and community GPS recording.

Treadmill walking distance is known to have good test-retest reliability, but exact protocols can vary between institutions and its results do not always correlate well with patient perception of their degree of limitation (Deen et al. 2000). These differences may in part be explained by the difference between capacity and performance, the former representing an individuals ability to perform in a controlled setting, and the latter performed by the individual in a day to day basis in the context of their own life (Organization 2001). These are known to be related, but measurably different, in individuals with lumbar spinal stenosis (Conway et al. 2011) and there is some evidence that self reported maximum walking distance has a stronger relationship with GPS recorded community walking distance than laboratory based measurements (Tew et al. 2013). Aside from Lau et al. (2017), none of the included papers attempted to measure objective walking performance, as opposed to capacity, despite its likely greater importance to patients quality of life.

Despite the above, the correlations from the papers using patient self-report were small, and hence, if a true underlying relationship exists between maximum walking performance and canal size it is likely to be weak.

In addition to walking ability, the included papers used in total 11 separate quality of life/disability measurement tools, none of which demonstrated a convincing and repeated relationship with the DS-CSA. In particular, the Oswestry Disability index, several SF-36 domains and VAS/NRS measurement of pain, were repeatedly assessed by multiple papers, consistently finding no significant relationship between the DS-CSA.

The included papers had several fairly consistent design flaws. A significant proportion of the included studies involved surgical or secondary care populations and, either implicitly or explicitly, used radiological criteria to include or exclude patients. While this is not surprising, given the consensus definition of symptomatic lumbar spinal stenosis includes both clinical and radiological criteria (Kreiner et al. 2013), it poses significant problems in studies designed to assess the relationship between radiological and clinical severity. Use of such criteria is likely to cause omission of patients with more mild presentations, both radiologically and clinically for surgical cohorts, and hence may reduce the papers power to detect a

correlation. Similarly, only a single paper made any attempt to control for confounding variables and all included correlation coefficients were bi-variate analyses. With such such an approach it seems equally possible that the papers finding correlations between DS-CSA and outcome variables were doing so because of some confounding factor, for instance age or gender, known to be related to the degree of narrowing (Ishimoto et al. 2012), or that the presence of patient factors, for instance pain sensitivity (H.-J. Kim et al. 2015), were masking a true underlying relationship.

This review has several limitations. For several of the models presented the statistical heterogeneity was high, and this is perhaps unexpected given the observational nature of the studies (Dwyer et al. 2001). Both Spearman and Pearson correlation coefficients have been pooled and, while these measure closely related but different attributes of the data, they are have been shown to approach each other and become similar with increasing sample sizes (de Winter et al. 2016). The meta-analysis presented by this review should be considered as an adjunct to the overall evidence rather than as being an authoritative estimate of the true underlying relationship or lack thereof.

2.5.2 Review 2

The aim of this review was to find radiological features associated with neurogenic claudication symptoms. While fourteen papers investigating MRI findings in both individuals with and without neurogenic claudication were found, all but one primarily investigated a case population with symptomatic lumbar spinal stenosis and hence included radiological criteria within their case, and sometimes control, definitions. While this approach has the advantage of conforming to the NASS guidelines and selecting cases with potentially surgically treatable disease, such an approach means any identified differences in the MRI studies are likely to be explained, at least in part, by the differences in radiological inclusion criteria rather than the differing symptoms or function between patient groups.

Only one included paper did not appear to separate participants on radiological features. Hamanishi et al. (1994) showed smaller DS-CSA measurements in neurogenic claudication patients, but performed the analysis on a level by level basis and did not seek to identify a threshold that separated the case and control groups. The paper also showed higher levels of multilevel stenosis (defined as a DS-CSA < 100 mm^2) within claudication patients compared to other lumbar conditions.

Neither this not the previous review includes evidence from myelography or CT based studies. This can be justified on the grounds that MRI has largely replaced these modalities as the primary investigation in suspected stenosis. While a recent diagnostic test accuracy review has shown MRI has an equivalent if not superior sensitivity and specificity to these modalities (de Schepper et al. 2013), its papers used anatomical rather than clinical reference standards.

In conclusion, while there are multiple imaging features that may potentially differentiate symptomatic lumbar spinal stenosis patients from other individuals undergoing MRI of the lumbar spine, the literature is insufficient to propose a clear definition of stenosis on the basis of its consistent associated with the key symptom of neurogenic claudication.

Chapter 3

Methodology

3.1 Overall aims

- To investigate the inter- and intra-rater reliability of a range of commonly used MRI based measurements described within the literature that measure structures potentially related to the diagnosis of anatomical lumbar spinal stenosis.
- To investigate how both qualitative and quantitative measurements of the different lumbar canal structures interrelate, how these measurements change with increasing age and other demographic factors and how these relationships change in patients with and without neurogenic claudication.
- To look for a threshold level of narrowing of the spinal canal or other radiological sign identified on MRI examinations of the lumbar spine that differentiate patients with and without neurogenic claudication and to assess the associated accuracy for doing so.
- To investigate the relationship between severity of clinical presentation and the degree of lumbar canal narrowing in a population with neurogenic claudication symptoms who have not been selected for inclusion based upon any anatomical criteria related to the lumbar canal.

The majority of the MRI data analysed in this thesis comes from participants of the BOOST RCT, whose methodology is presented in the next section.

The presented methodology aims to avoid some of the problems identified in the current literature and discussed previously in Chapters 1 and Chapter 2. Particularly, by focusing on patients with neurogenic claudication rather than the full diagnosis of symptomatic lumbar spinal stenosis (i.e. not specifying a degree of canal narrowing required to be included in any case population) underlying assumptions about the anatomical causes of neurogenic claudication are avoided, potentially less severe presentations that could be excluded on radiological grounds will still be included, and the proposed case-control type study can be more sure that any differences in imaging between the two participant groups are due to the difference in symptomatic status rather than because of any differences in anatomical inclusion criteria for the two groups.

3.2 BOOST: better outcomes for older adults with spinal troubles — a randomised controlled trial

Note: the full protocol of the BOOST RCT has been published by Williamson et al. (2018).

3.2.1 Rationale

Current management options for patients with NC diverge depending on the severity of symptoms, their duration, and the presence of anatomical stenosis as demonstrated on imaging (Kalff et al. 2013). For patients with both neurogenic claudication and radiological stenosis (i.e. symptomatic LSS), surgical relief of the stenosis can be undertaken, but this is generally only considered after a period of conservative management (Machado et al. 2016). Surgery is also expensive, carries a risk of complications including infection and CSF leak (Deyo et al. 2010), and is of uncertain efficacy — a 2016 Cochrane society review concluded with a statement indicating the authors had little confidence that the available evidence base showed a conclusive benefit from surgical treatment when compared to conservative management options (Zaina et al. 2016). Surgery also showed a complication rate between 10% to 24% in contrasted to the low morbidity associated with non-surgical management.

Non-surgical treatment generally involves physiotherapy alongside other management options including facet joint injections and oral analgesia (Ammendolia et al. 2013). Patients without clear anatomical changes to explain their NC symptoms are dependent upon these non-operative treatments given the lack of a surgical target for intervention. The

evidence base for non-operative care for neurogenic claudication patients is of low quality: the available clinical trails suffer from significant limitations in follow up length and have generally small sample sizes (Ammendolia et al. 2013). Physiotherapy is a commonly used part of this non-operative management (Comer et al. 2009) but there is little to guide both what the best form of physiotherapy intervention is and what its likely efficacy is in relief of symptoms and improvement of disability (Macedo et al. 2013a).

The BOOST multicenter RCT is designed to address this gap in the evidence base, aiming to develop and test the efficacy of a physiotherapy intervention in patients with NC. As part of its design it will also explore potential predictors of any identified treatment effect, as well as general predictors of patient prognosis over the course of the trial and it's follow-up. Among the potential predictors explored will be a set of MRI based measurements of spinal canal size.

3.2.2 Objectives

- To estimate the clinical and cost-effectiveness of a physiotherapist delivered physical and psychological intervention for older adults with neurogenic claudication when compared to current best practice.
- To explore whether radiological measurements of canal size, indicators of frailty or behavioural factors can identify groups of participants more like to benefit or fail to benefit from the intervention using a pre-specified subgroup analysis.
- To conduct in parallel with the main RCT a longitudinal qualitative study to collate participant experiences of living with neurogenic claudication.

3.2.3 Overview of study design

The overall study design is that of a multicenter RCT.

Community dwelling participants with neurogenic claudication were recruited at multiple centres spread throughout the UK, including both secondary care and primary care settings. On recruitment, participants were randomised in a 2:1 ratio to either the BOOST physiotherapy intervention or to current physiotherapy based best practice.

The BOOST intervention consistent of physiotherapist delivered combined physical and psychological program. Each participant first attended an individual session followed by 12 group sessions, each involving 6 participants, delivered over a 12 week period. During the individual sessions participants underwent an assessment followed by a tailored prescription of a series of exercises designed to target muscle strength, flexibility and balance. The group sessions consisted of education and group discussion, followed by warm-up and circuit exercises, followed finally by a walking circuit, the latter aimed at increasing mobility and walking self-efficacy. The education and psychological component focused on cognitive behavioural techniques, pain management strategies and encouragement to engage with the home exercise regime. On conclusion of the group exercises two follow-up phone calls were made to encourage ongoing adherence to the home exercise program Ward et al. (2019).

The comparator best practice arm was given advice by a physiotherapist ideally in a single session, with up to two review sessions as needed.

Participants were followed up for 12 months post recruitment with the primary assessed outcome being the ODI, a measure of back pain related disability.

3.2.4 Recruitment

BOOST began recruitment in July 2016, eventually opening in 15 centres across the UK. Recruitment was closed in August 2019.

Potential participants were identified through one of two pathways. The first recruited from secondary care physiotherapy and spinal clinics. Clinic staff provided potentially eligible patients with information on the study and also gained consent for further contact by research staff. The second recruited participants from a related primary care population based cohort study — the Oxford pain, activity and lifestyle cohort study (OPAL). OPAL is currently ongoing at 34 primary care centres in the UK and has recruited participants from a random sample of the individuals registered at each centre. The initial OPAL study questionnaire includes a set of questions to identify participants who would be potentially eligible for BOOST and such participants were approached for BOOST recruitment.

Overall 884 potential participants underwent telephone screening for inclusion, 732 identified by clinical staff in secondary care clinics and 152 recruited from the 5409 OPAL cohort study participants. After telephone screening 360 were excluded, with 123 not being meeting eligibility criteria, 166 declining consent and 71 whom it was not possible to contact. 524 underwent a research clinic assessment with a total of 438 recruited. Of those screened at the research clinic, 28 were found to not meet eligibility criteria, 1 declined consent and 57 failed to attend the appointment.

At the time of writing (Jan 2020) 53 participants have withdrawn from the study (12.1%), and 4 participants have died from events unrelated to the RCT. Of those withdrawing, 3 withdrew consent for ongoing use of their data and hence are absent from any further analyses.

3.2.5 Participant eligibility

For the purposes of the BOOST trial neurogenic claudication was defined using a set of self report questions which have been found to have excellent sensitivity and specificity for identifying NC (de Schepper et al. 2013). The questions are listed below.

Primary question:

- In the past 6 weeks, have you had back pain and/or pain or other symptoms such as tingling, numbness or heaviness that travelled from your back into your buttocks or legs?

Secondary questions:

- Does standing make the pain or symptoms in your buttocks or legs worse?
- Does walking make the pain or symptoms in your buttocks or legs worse?
- Does sitting down make the pain or symptoms in your buttocks or legs better?
- Does bending forward (e.g. to push a shopping trolley) make the pain or symptoms in your buttocks or legs better?

In order to be judged as having neurogenic claudication participants had to answer the primary question as yes and also answer at least one secondary question as yes.

In addition, potential participants were also required to fulfil the eligibility criteria listed below. The exclusion criteria primarily relate to fitness to participate in the BOOST intervention.

Each potential participant was initially screened for eligibility by a BOOST researcher first by telephone call and then by attendance at a research clinic appointment. The clinic appointment involved ensuring the symptoms were consistent with NC, screening for cauda equina syndrome or other serious pathology requiring further investigation, an abbreviated mental test (AMT) test and a mobility assessment. For those meeting all eligibility criteria, written informed consent was then taken followed by baseline data collection.

Participant inclusion criteria:

- Registered with a primary care practice
- 65 years and over
- Participant is willing and able to give informed consent for participation in the RCT
- Reports symptoms consistent with neurogenic claudication

Participant exclusion criteria:

- Living in a residential care or nursing home
- Has a terminal condition with a life expectancy of less than 6 months
- Any substantial health or social concern that, in the opinion of the patient's general practitioner (GP), would place the patient at increased risk or inability to participate including known inability to provide informed consent (e.g. dementia)
- Unable to walk 3 meters without the help of another person
- On a surgical waiting list
- Presents with cauda equina syndrome or signs of serious pathology requiring immediate referral for investigations
- Cognitive impairment (defined as an Abbreviated Mental Test score of 6 or less)
- Registered blind
- Unable to follow verbal instructions which would make participation in the exercise group impractical; including severe hearing impairment not corrected by a hearing aid or inability to follow safety instructions (e.g. English comprehension)

3.2.6 Randomisation

Randomisation into the BOOST intervention and control groups occurred following baseline data collection and used a UK Clinical Research Collaboration approved web-based service. The recruiting researcher, those performing follow-up data collection and the imaging team are blinded to treatment allocation. Participants and the treating physiotherapists cannot be blinded to treatment allocation but are asked not to share their treatment allocation with researchers.

Randomisation was stratified by centre, gender and age in a 2:1 ratio of BOOST intervention to control group participants. Of the 438 included, 295 were randomised to the group physiotherapy program and 143 to best practice advice.

3.2.7 Data collection

The full outcome measures for the trial and baseline data points are listed in Appendix C.3. Data on all outcome measurements is collected at initial recruitment and both 6 and 12 months post randomisation. Four of the clinical outcomes are of particular importance for this thesis:

The change in Oswestry disability index (ODI) v 2.1a from baseline at 12 months post randomisation is the primary outcome for the BOOST RCT. The ODI is a widely used, self-completed measurement of low back pain rated disability and function, and ranges between 0 and 100 with higher numbers representing greater disability (Fairbank et al. 2000). The index consists of 10 items, covering topics that assess how back or leg pain is affecting the patient's ability to manage everyday life, including overall pain intensity and restriction to walking ability. For the purposes of BOOST, a question on sexual function was omitted. Each item is scored from 0 – 5, with the percentage of total possible score taken as the final index. A change of 10 points in the ODI is required to be clinically meaningful with smaller changes potentially attributable to measurement error.

The Zurich claudication questionnaire (ZCQ) is a self-report outcome which quantifies severity of symptoms and physical function over the past month alongside patients satisfaction with treatment (Stucki et al. 1996). It is also often referred to as the Swiss spinal stenosis questionnaire in the literature. Only the ZCQ symptom subscale was used for

BOOST, which itself is subdivided into a pain domain (consisting of 4 items) and a neuro-ischemic domain (3 items). Each item is rated on a score from 1 to 5. The final score for subscale and each domain is found by averaging its associated items and then expressing the average as a percentage of the highest possible score. Increasing values represent worsening symptoms.

The EuroQol 5-dimensions (EQ-5D) — specifically the EQ-5D-5L — is a instrument for measuring quality of life/health status (EuroQol Group 1990). The system consists of five dimensions — mobility, self-care, usual activities, pain/discomfort and anxiety/depression — each with 5 levels coded from 1 to 5; with increasing score representing greater disability. The results of the EQ-5D across all dimensions can be expressed as a single index number by using a set of values (weights) that grade the degree of compromise the score for each dimension would have on overall quality of life. The set of weights used is country specific and non-linear in nature, found by polling a representative sample of individuals from that country. The index ranges between 0 and 1 (expressed as a percentage for the purpose of this thesis) with higher values representing higher quality of life. The weight set used within this thesis is described by Deyo et al. (1998).

The 6 minute walk test (6-MWT) measures the distance an individual is able to walk in 6 minutes on a flat indoor hard surface (Bennell et al. 2011). Rest is allowed as required. In addition to the total distance walked, researches also assessed whether individuals were symptomatic before starting the test, whether they developed claudication symptoms during the test and if so at what distance.

The other secondary outcomes for BOOST cover a range of self-report items covering both treatment targets: strength, balance, mobility, frailty, etc; but also cognitive and behavioural factors that potentially mediate the efficacy of the intervention. The most important of these include the Tilburg frailty index (TFI), a measure of frailty; the STarT back score, a primary care screening tool that aims to categorise individuals as low, medium or high risk for developing persistent disabling symptoms, and the short physical performance battery (SPPB), a physiotherapist administered tool for assessment of lower leg function which includes assessment of standing, walking and chair rising ability.

3.2.8 BOOST MRI imaging

Note: the full standard operating procedure (SOP) for the imaging section of the BOOST randomised controlled trial is presented in Appendix C.1

As part of baseline data collection, MRI studies of the lumbar spine were collected for as many BOOST participants as possible. Where a participant had a pre-existing MRI study of the lumbar spine performed in the 12 months before randomisation this was collected by electronic transfer from the performing institution. Where a participant did not have a pre-existing MRI study within the correct time frame, or where it did not prove possible to collect the pre-existing MRI study, patients were asked to consent for a new MRI study specifically for the purposes of the trial. These MRIs were performed either at the recruiting centre, or where this was not possible, by referral for a private study at an outside institution.

Imaging protocol

MRI lumbar spine protocols are fairly standardised across NHS radiology departments. The MRI studies are performed supine with T1 and T2 weighted sagittal imaging and T2 weighted axial imaging of at least the lower three discs performed. Additional imaging at other levels is often added at the discretion of the performing radiographer on the basis of their interpretation of the sagittal imaging. The axial imaging will either be taken as separate blocks, each angulated parallel to and centred on the discs, or as a single block extending from L3 to S1. The former was preferred for the purposes of the study and from the authors experience is the most common technique in NHS hospitals. The imaging parameters were expected to be similar to table 3.1.

Where studies were performed specifically for the purposes of the trial they were asked to be conform with the above scanning protocol, with angled axial blocks. When collecting pre-existing MRIs non-standard sequences were accepted where they allowed collection of data from the MRI in an equivalent manner to the standard sequences. For instance, volumetric T2 sequences were accepted as they could be reformatted to produce images similar to a standard T2 axial sequence.

All BOOST MRIs performed for the trial had a standard radiology report produced, which, on request, was provided to the study participant at the end of the trial. Except where necessitated by imaging findings requiring urgent clinical action, researchers other than the trial imaging team were blinded to MRI study results.

MRI collection

Overall, 354 MRI studies were available at conclusion of recruitment (80.8% of all participants). Of these 179 (50.6%) were pre-existing clinical MRIs and 175 (49.4%) were performed specifically for the purposes of the trial.

Participants did not have MRI studies included in the study for the following reasons:

- 4 participants reporting a pre-existing MRI:
  - 3 – No study at site reported by participant
  - 1 – Private study unavailable for transfer
- 77 participants reporting no previous lumbar spine MRI:
  - 69 – Declined consent or had contraindications for MRI

Table 3.1: Expected MRI parameters for BOOST MRI studies

Sequence	FOV	Slice	Gap	TR	TE	ETL	Phase	Freq	NEX
T2 sagittal	370	13/4	1	4061	102	23	320	512	3
T1 sagittal	370	13/4	1	446	11	3	224	416	3
T2 axial	200	30/4	1	4955	111	25	224	320	3

- 7 – Recruiting centre unable to provide MRI
- 1 – Unable to contact participant for trial MRI referral

Multiple BOOST participants reported having a pre-existing clinical MRI study that subsequently proved impossible to include. These participants generally misremembered where the study had been performed, making transfer of the images impossible, or were found to have a pre-existing study of the wrong modality or body part. Where this occurred, or the pre-existing lumbar spine MRI had been performed greater than one year post recruitment, the participant was generally offered a new MRI study specifically for the purposes of the trial. This was not possible in 10 participants, 6 of whom had MRI studies older than the 1 year limit (range 378 to 485 days before recruitment) and 4 participants without any pre-existing study. A decision was made to include the older MRI studies on the grounds that they did not exceed the 1 year limit by a large enough time period that a significant change in the patients anatomy would be anticipated. In total, 77 participants reported no previous MRI study but did not have a MRI study for the purposes of the trial. The majority of these participants either declined consent for an MRI study or had contraindications to MRI scanning.

The median absolute time from recruitment to the MRI study was 78 days; 86 days for those with a pre-existing MRI study and 55 days for those having an MRI study for the purposes of the trial. The overall range of duration from recruitment to MRI study extended from 485 days from before recruitment (the oldest of the 6 MRI studies mentioned above) to 678 days post recruitment (figure 3.1). Overall 28 participants had an MRI more than one year after their initial recruitment date, this mostly occurred due to organisation difficulties in arranging the MRI study. These delayed MRI studies were included, again on the expectation that no significant change in the patients anatomy over the time periods involved is expected.

3.2.9 Study interventions

Experimental intervention: the BOOST program

A detailed description of the BOOST intervention for older adults with neurogenic claudication has been published by Ward et al. (2019).

The experimental intervention consisted of an individual assessment and home exercise prescription followed by twelve 90 minute group sessions conducted over a period of 12 weeks. The sessions begin twice weekly, reducing in frequency over the course of the program. In general one physiotherapist delivers each session to groups of six participants.

The format of each session is standardised, beginning with an education section (30 minutes) followed by an exercise section (60 minutes). The education section covers behavioural change strategies designed to increase adherence with the home exercises. The exercise section starts with warm-up seated exercises (arm raises, knee lifts, trunk and pelvic rotation) followed by a circuit of strengthening, stretching and balance exercises including sitting knee extension, sit to stand, standing hip abduction and standing hip extension. The session closes with a walking circuit targeting walking ability and fitness. The exercises are tailed to each participant and progressed over the twelve sessions.

Participants are asked to complete their home exercise routine, prescribed in the initial session, at least twice a week. At the conclusion of the group sessions the physiotherapist carries out two follow-up telephone reviews with each participant to promote ongoing compliance.

Control intervention: best practice advice

The comparator intervention for the BOOST program is best practice advice delivered by a physiotherapist, as advised by a survey of current practice (Comer et al. 2009).

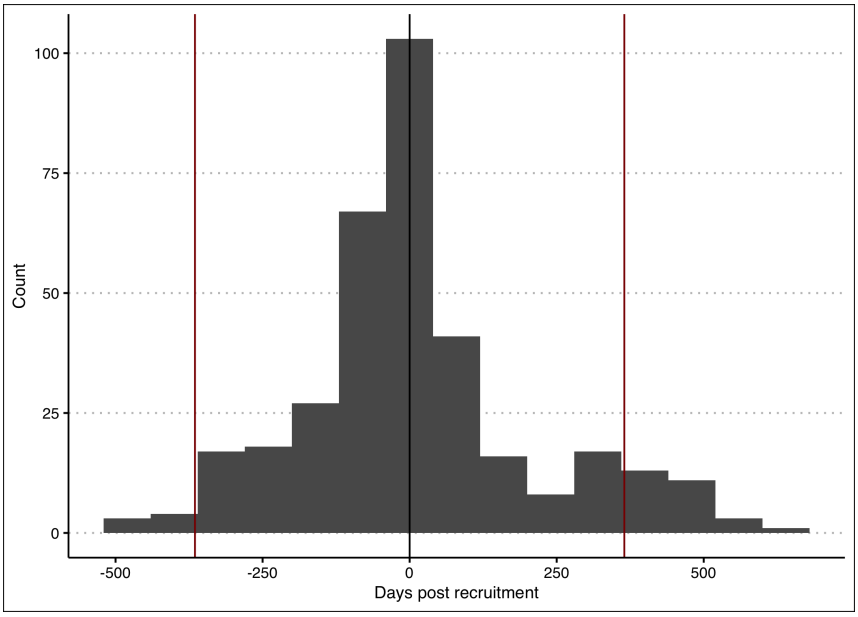
Participants attend an initial appointment of up to 1 hour consisting of an assessment followed by advice, education and prescription of a home exercise programme. The latter usually consists of flexion and trunk stabilisation exercises but can also include other exercises based on the participant's assessment. If required a walking aid may be prescribed. While a single session is preferred, a maximum of two half-hour review appointments is also permitted to review advice, exercises and walking aid use as needed.

3.2.10 Planned sample size and data analysis

The required BOOST sample size was calculated at 402 participants. In order to provide an 80% power, with 0.05 alpha, to detect a 5 point difference in ODI (with standard deviation (SD) of 15) a sample of 321 participants was required. An additional 20% was then added to account for potential participant loss to follow-up by the conclusion of the trial.

The analysis of treatment effect will be done on "intention to treat" principles. The primary outcome

Figure 3.1: Histogram of BOOST participant MRI study timing relative to recruitment date. Red lines mark one year pre- and post-recruitment.



will be analysed using a linear multi-variable regression, controlling for region, therapist effect and other identified prognostic variables. A secondary analysis will focus on the effect of adherence with intervention.

Multiple *a priori* subgroup analyses of treatment effect size were planned exploring the effect of age, gender, frailty, pre-existing physical performance and a set of MRI defined subgroups.

3.3 Imaging control group

Note: The full protocol for the recruitment of the imaging control group is included in Appendix C.2

An imaging control group without neurogenic claudication symptoms was recruited prospectively for comparison with the BOOST participants. The original BOOST ethics approval did not include an imaging control group. Additional research ethics committee and health research authority approval for this study was therefore sought and granted (REC reference: 18/LO/1367).

Consecutive GP referrals to the University Hospital Coventry and Warwickshire radiology department for lumbar spine MRI for investigation of low back pain were screened for eligibility for the imaging control group through the clinical information provided with the referral. Potentially suitable patients were mailed a invitation letter and participant information sheet, along with their booking information.

On arrival in the radiology department, patients who indicated they would be willing to participate in the research study underwent screening for eligibility by a researcher. Those who answered the first question from the BOOST definition of neurogenic claudication as yes (i.e. those who reported a history of pain, discomfort or dysaesthesia radiating from the back into the buttocks or legs, see § 3.2.5) were excluded. Such an approach will exclude some individuals who do not have neurogenic claudication, including those with non-claudicant lower limb radicular pain, but was thought necessary as the definition of symptomatic lumbar stenosis specified by the NASS definition of symptomatic LSS is not universally accepted and some practitioners consider radicular pain without claudication to form part of the spectrum of the disease (see figure 2.3). The other inclusion and exclusion criteria for the control group mirror those of the BOOST study to ensure the two groups are otherwise comparable. Consistent with this, all control participants were over 65 years of age.

Patients meeting the eligibility criteria were asked to give written informed consent for the study and an MRI was performed, with protocol as that described in § 3.2.8.

The initial aim was to recruit 150 control participants (a 2 to 1 ratio with BOOST participants, assuming MRIs would be available for 300 BOOST participants at close of recruitment). This sample size, with an alpha of 0.05, gives beyond a 80% power to detect a difference of 10 mm^2 DS-CSA between symptomatic and control groups, assuming a standard deviation of 40 mm^2, the latter estimated from the findings of Hamanishi et al. (1994). The planned sample size also gives beyond a 95% power to detect an area of 0.6 under a receiver operator characteristic (ROC) curve (Obuchowski 2005). The planned sample size

was over-powered to aid a planned computer-vision based analysis of the MRI data. No accepted way to calculate required sample sizes for such an a approach exists but larger samples are generally thought to improve the chances of success.

Recruitment of low back pain participants without NC to act as an imaging control group began in October 2018. Recruitment closed in August 2019 after recruitment of 63 participants, falling short of the original target sample size. During the recruitment period a far higher than expected number of potential participants screened were excluded due to reporting leg pain in the previous 6 weeks, substantially slowing recruitment. A small number of those screened (5 patients) declined participation. No participant was excluded for any other inclusion or exclusion criteria.

The decision to halt control recruitment short of the targeted participant number was made due to reaching the end of the funding period for the PhD and the fact that a satisfactory predicted power had been reached for the purposes of the case-control study — 63 control participants compared with a subgroup of 273 BOOST participants would give an 80% power to detect an area under the ROC curve of 0.6 with an type 1 error risk of 0.05.

3.4 Participant demographic data

The demographic data for both case and control populations is presented in table 3.2. No significant difference between age, body mass index (BMI) or gender distribution was identified between the BOOST and imaging control group participants. The mean height of BOOST participants was 3 cm shorter than those of imaging controls (t-test, $p = 0.01$).

The ethnic makeup of the BOOST and imaging control participant groups is shown in table 3.3. The BOOST participant group was significantly different to the imaging control group in terms of ethnic make-up (Chi-squared, $p < 0.01$) included a larger number of ethnic groups (9 as opposed to 5) but an overall higher number of white British individuals compared to the imaging control group.

No significant difference was demonstrated between any demographic data collected when comparing BOOST participants where an MRI was available to those without an MRI study at the conclusion of data collection. There was a significant difference in ethnicity make-up of the BOOST participants with a pre-existing clinical MRI study compared to those having an MRI for the purposes of the trial (Chi-squared, $p = 0.01$), with a lower percentage of white British individuals in the group without pre-existing imaging. No other collected demographic data significantly between these groups.

There were multiple weak relationships between demographic variables and clinical severity variables in the BOOST participants — see figure 3.2. Age was weakly negatively correlated with both total 6 minute walking distance ($R = -0.14$, $p = 0.006$) and the SPPB score ($R = -0.11$, $p = 0.03$). BMI was positively correlated with the ODI ($R = 0.20$, $p < 0.001$) and negatively correlated with total walking distance ($R = -0.24$, $p < 0.001$), SPPB score ($R = -0.25$, $p < 0.001$) and EQ-5D index ($R = -0.12$, $p = 0.03$). BMI also significantly differed between STarT back score classes (Kruskal-Wallis, $p = 0.02$) with significantly higher mean BMI in medium and high risk groups compared to the low risk group (Dunn's test, $p = 0.01$ for both comparisons). Similarly, height was positively correlated with total walking distance ($R = 0.11$, $p = 0.04$) and negatively correlated with frailty ($R = -0.12$, $p = 0.03$).

Women had lower performance on the SPPB and a higher level of frailty (TFI) compared to men (Man-Whitney, $p = 0.03$ and $p < 0.001$ respectively). There were no other significant differences in clinical severity variables between genders. Those BOOST participants who had a pre-existing scan within the previous 12 months from recruitment unsurprisingly reported significantly more severe neurological symptoms (i.e. Zurich claudication questionnaire neuroischaemic domain (ZCQ NI), Man-Whitney, $p = 0.005$) compared to those who did not have a recent MRI study available. No other differences in clinical severity scores were identified between these groups. Ethnicity was significantly related to frailty scores (Kruskal-Wallis, $p < 0.001$), but no other clinical severity variable.

3.5 MRI assessment

During the original design of the BOOST RCT, 3 MRI based measurements were pre-specified for use as grouping variables for the planned sub-group analysis of BOOST treatment effect: the dural-sac cross-sectional area (DS-CSA), the lateral recess depth (LRD) and the neural exit foramen diameter (NFD). These measurements were chosen on the basis of their common usage in the literature and their coverage of three main sites of potential cauda nerve root impingement within the spinal canal.

Deciding what, if any, additional measurements to perform on each of the collected MRI studies was made difficult by the large number of measurements of the lumbar canal described in the literature and the relative lack of available evidence to distinguish them based on their potential clinical relevance (see § 2.1). It was decided to assess a range of measurements for suitability starting with the meas-

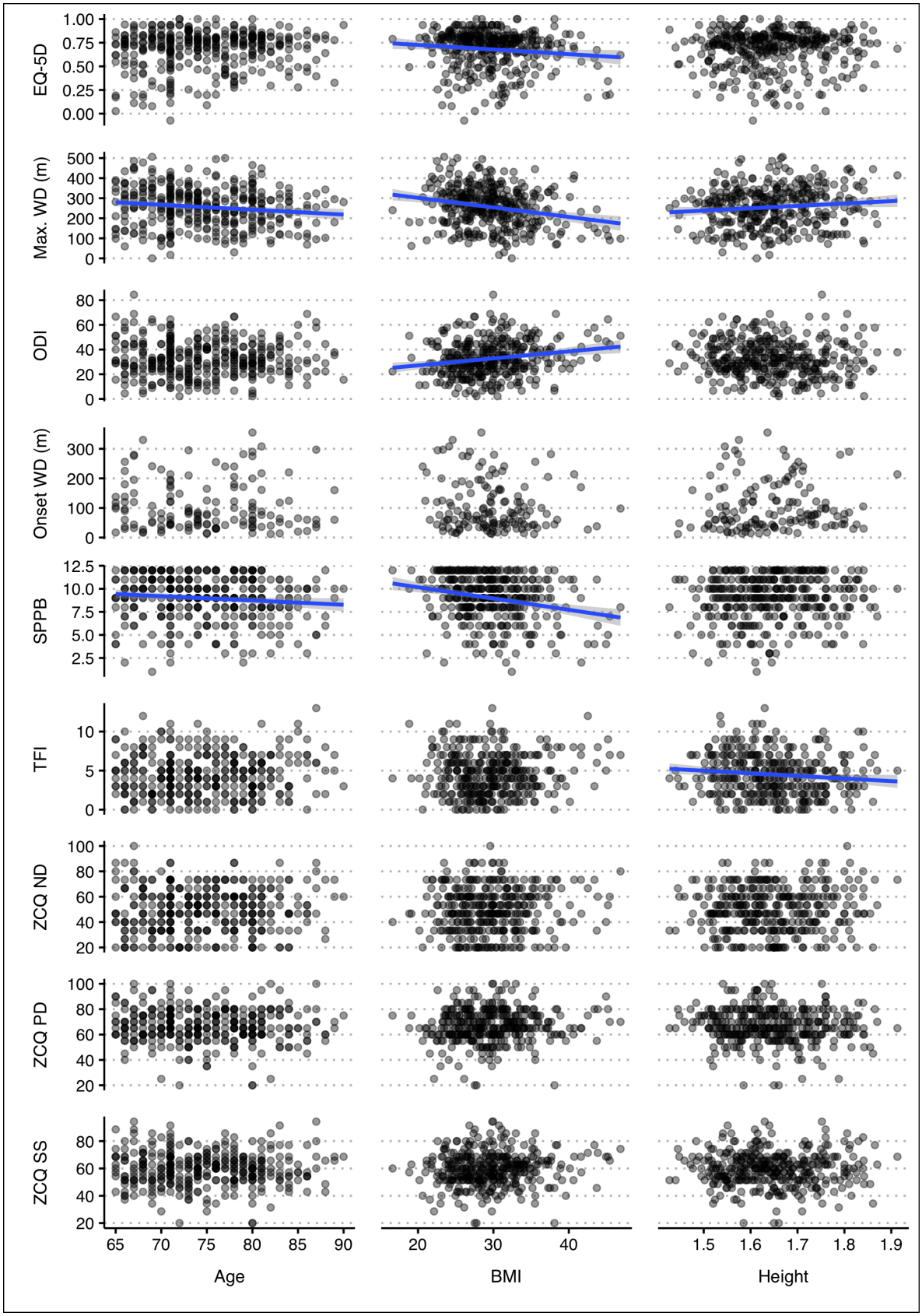


Figure 3.2: Scatter plots showing relationships between clinical severity scores and the age, BMI and height of participants. Onset walking distance (WD) refers to the walking distance to symptom onset. Lines of best fit and their confidence intervals are drawn for variables showing a significant Pearson's correlation.

Table 3.2: Demographic data for both imaging BOOST and imaging control participants. BOOST participants are further broken down into those with and without MRI scans, and for those with an MRI scan, those having one for the purposes of the trial vs a pre-existing clinical scan.

	Mean Age (years)	Mean Height (m)	Mean BMI (kg/m^2)	% Male	% White British
Imaging Control Participants	73.90	1.68	30.42	53.97	85.71
All BOOST Participants	74.81	1.65	29.93	43.45	88.51
BOOST Participants without MRI	75.00	1.66	30.90	49.38	88.89
BOOST Participants with MRI	74.77	1.65	29.71	42.09	88.42
BOOST Participants with Trial MRI	74.82	1.65	29.86	39.77	93.18
BOOST Participants with Clinical MRI	74.71	1.64	29.56	44.38	83.71
P – BOOST vs Imaging Control	0.29	0.01	0.59	0.13	< 0.01
P – BOOST: MRI vs no MRI	0.77	0.41	0.06	0.27	0.32
P – BOOST: Clinical vs Trial MRI	0.85	0.59	0.56	0.40	0.01

Table 3.3: Comparison of the ethnic makeup of BOOST participant and imaging control groups.

Ethnicity	BOOST Participants		Imaging Controls	
	N.	%	N.	%
Black or Black British	12	2.7	0	0.0
Indian	9	2.1	3	4.8
Mixed	3	0.7	0	0.0
Other	3	0.7	0	0.0
Other Asian	0	0.0	1	1.6
Pakistani	6	1.4	1	1.6
Prefer not to say	3	0.7	4	6.3
White British	385	87.9	54	85.7
White Other	15	3.4	0	0.0

urements suggested by Andreisek et al. (2014) as core reporting criteria for LSS (described further in § 1.5.4) and iteratively adding and removing measurements to deal with perceived shortcomings in measurement reliability, coverage of relevant anatomical structures and the overall time taken to assess a single MRI spine. Where necessary, the originally described measurement technique was modified to improve measurement performance, as described in subsequent chapters which discuss the characteristics of each measurement.

Four observers took part in this process, performing two rounds of measurement separated by several month on a set of 22 MRI studies randomly selected from all the MRIs collected for both BOOST and imaging control groups. The observers consisted of a radiology registrar in the fifth year of training (the thesis author), a neuroradiology fellow and two consultant radiologists, one with a sub-speciality MSK interest. Images were viewed using *Horos*, an open source DICOM viewer (Project 2019).

Each observer was trained in performing the measurements by example, followed by performing a set of measurements under observation, with further discussion as required. The measurements were entered into an electronic form which also gave written and illustrated guidance on how to perform each measurement as it was performed. Individual measurements that were markedly discrepant between observers were manually inspected, and where a clear error in data entry has occurred, such as a misplaced decimal point, the error was corrected before further data analysis.

For each proposed measurement the inter-rater reliability was calculated using intra-class correlation (ICC) for quantitative measurements and both Fleiss' kappa statistic and percentage agreement scores for qualitative measurements. For calculation of the kappa statistic disagreements were weighted according to their squared distance from perfect agreement. Each statistic was calculated for each possible pairing of observers and a mean was calculated, weighted by the number of shared measurements for each observer pair.

In considering whether to include a measurement in the final data-set the inter-rater reliability, degree of additional information provided by the measurement compared to other measurements assessing the same anatomical structure and ease of measurement were considered.

The measurements selected for use on the full MRI data-set are presented in tables 3.4 and 3.5. These measurements were performed on all MRI studies by a single observer (the thesis author) using the same electronic form setup as was used in the initial measurement assessment phase. After a gap of several months a group of 32 MRIs were reassessed to allow calculation of intra-rater reliability.

MRI observers were blinded to imaging control or BOOST participant group status, BOOST treatment allocation and baseline symptoms. Imaging control vs BOOST participant blinding was achieved by

Table 3.4: Quantitative MRI measurement reliability.

Variable	Inter-rater ICC	Intra-rater ICC
Dural-sac cross-sectional area (DS-CSA)	0.92	0.97
Lateral recess depth (LRD)	0.65	0.87
Neural exit foramen diameter (NFD)	0.50	0.67
AP canal diameter (APD)	0.49	0.85

Table 3.5: Qualitative MRI measurement reliability.

Variable	Inter-rater		Intra-rater	
	κ	% agreement	κ	% agreement
Presence of Scoliosis	1.00	100.00	1.00	100.00
Spondylolisthesis Grade	0.76	96.30	0.80	94.90
Lee Grade	0.68	87.10	0.68	92.40
Presence of Disc Herniation	0.28	85.60	0.71	96.20
Schizas Grade	0.77	83.80	0.91	93.70
Bartynski Grade	0.70	71.00	0.82	73.10

automated assignment of new unique and randomised identification numbers to each study, with modification of the DICOM headers to remove any potentially identifying information other than the new identification (ID). This automated assignment and header modification was performed in blocks by a custom script written in Python 3 and run by an individual independent of the research team. The key, liking new study ID to BOOST or control group ID, was made available to the thesis author at the completion of the grading processes to allow unblinded assessment of the data. Blinding with regard to BOOST treatment allocation was controlled by the central BOOST research team with only the BOOST statistician, who performed the main trial and subgroup analysis, unblinded to treatment allocation.

3.6 Statistical analysis

All statistical analysis was performed in the R environment for statistical computing (R Core Team 2020).

In general, quantitative measurements were first assessed for a normal distribution by inspection of quantile-quantile plots and the Shapiro-Wilk test. Statistical testing for for a difference in distribution between two groups was either performed by t-test, where the distributions could be expected to be normal, or by Man-Whitney U test, when a non-parametric test was required. Similarly, for comparison of more than two distributions, one-way analysis of variance (ANOVA) or the non-parametric Kruskal-Wallis one-way analysis of variance was used with pair-wise t tests or Dunn's test for analysis of specific sample pairs respectively.

The statistical significance of the difference in distribution of categorical variables between groups was assessed using the Chi-squared test.

Assessment of correlation between pairs of quantitative variables is assessed by Pearson correlation coefficient and for groups of variables by multivariate linear regression. R^2 values are calculated to aid interpretation where needed.

The ability of different measurements to separate BOOST participants and the imaging control groups was performed by ROC analysis. The area under the curve (AUC) was calculated for each ROC curve with comparison between AUCs by DeLong's test. The "best" threshold measurement for separation of the two imaging groups for each analysed measurement was found by finding the point on the ROC curve with the greatest Yoiuden's J statistic.

Where the statistical approach deviates from the above it is explained alongside the presentation of the relevant results. This is particularly the case in Chapter 8 where multiple machine learning algorithms were trained with the goal of separating NC and LBP with greater accuracy than that possible using single measurements of the canal explored in earlier chapters. An explanation of machine learning approaches in general, alongside the exact techniques and algorithms used is presented within Chapter 8.

3.7 Chapter summary

This chapter set out the methodology for for a diagnostic cross-sectional study based upon data collected from participants of the BOOST RCT alongside a separately recruited control group. The methodology relating to participant recruitment, clinical data collection and MRI acquisition, and interpretation were set out.

Minor differences were found in the demographic factors between the two participant groups, suggesting a degree of sampling error, and there were also weak correlations between some clinical severity scores and demographic factors. These relationships should be borne in mind when interpreting results in subsequent chapters.

The results from the cross-sectional study are presented in Chapters 4 – 8.

Chapter 4

The Central Canal

Note: as described in Chapter 3 BOOST participants were recruited on the basis of their presentation with neurogenic claudication (NC) symptoms, while the imaging control group had non-specific lower back pain (LBP) without features of NC or radiculopathy. To emphasise the difference in group symptomatology, these participant groups will be referred to as NC and LBP groups respectively during this and subsequent chapters.

4.1 Measurement method and reliability

Dural-sac cross-sectional area (DS-CSA)

The DS-CSA was included in Andreisek's core reporting criteria as a research use only measurement and was the measurement of central canal stenosis prespecified for the BOOST sub-group analysis of treatment effect.

MRI observers were asked to select the axial slice for each lumbar spinal level that both clearly passed through the posterior disc-annulus and showed the most severe degree of dural sac narrowing. Where there were several slices passing through the posterior annulus with a similar degree of narrowing, the slice showing the minimal visible amount of vertebral body and maximum coverage of the lamina by the ligamentum flavum was selected — i.e. the slice closest to the mid disc level. The dural-sac area was then measured by drawing its borders using a closed polygon tool provided by *Horos* (as per figure 1.5 I).

The inter- and intra-observer reliability of the DS-CSA was excellent with ICC of 0.92 and 0.97 respectively.

Inspection of MRI images from spinal levels where measurements were markedly discrepant between observers showed a number of common themes. Typically such levels had larger amounts of epidural fat visible that obscured the boundary between dural sac and epidural fat due to its similar signal intensity on T2 weighted sequences to CSF. Axial T2 sequences were chosen over T1 for the measurement, despite the latter's better ability to demonstrate the interface between epidural fat and CSF, due to its much more common availability among the collected MRI studies. Another common issue occurred in severely stenotic levels where there was minimal visible CSF or epidural fat, making separation of the dark compressed nerve roots within the dural sac from the dark posterior disc annulus and ligamentum flavum difficult. Despite these issues on a small number of slices, the measurement reliability remained excellent.

The distribution of DS-CSA measurements diverged significantly from a normal distribution (Shapiro-Wilk $W = 0.98$, $p < 0.001$). The divergence occurred at both the upper and lower tails of the distribution (see figure 4.1).

AP Canal Diameter (APD)

The APD was included as a measurement due to its historical importance and potential use as a marker of developmental canal size, as discussed in § 1.4.1. It was measured between the anterior cortex of the lamina anteriorly to the posterior vertebral body cortex on mid-sagittal T2 weighted images. The measurement was made parallel to a line drawn through the mid-vertebral body (see figure 1.4, measurement 10).

The inter-rater reliability of this measurement method was found to be moderate with ICC of 0.49. It showed a good intra-rater reliability with ICC of 0.85.

On discussion between observers, difficulties in measurement contributing to the comparatively low inter-rater reliability likely involved difficulties in choice of mid-sagittal slice, particularly in studies with a significant scoliosis and difficulty in determining the exact location of vertebral body and lamina cortex given their close relationship to other structures (e.g./ the ligamentum flavum) which also appear dark on T2 weighted imaging. The basivertebral vein occasionally caused difficulties by entering the vertebral body in the plane of the measurement obscuring the posterior vertebral body cortex.

The distribution of the APD measurements were also found to be non-normal (Shapiro-Wilk $W = 0.95$, $p < 0.001$) with the divergence from normality occurring at the upper tail of the distribution (figure 4.1). In spondylolysis, bilateral breaks in the pars interarticularis occur, whose aetiology is uncertain

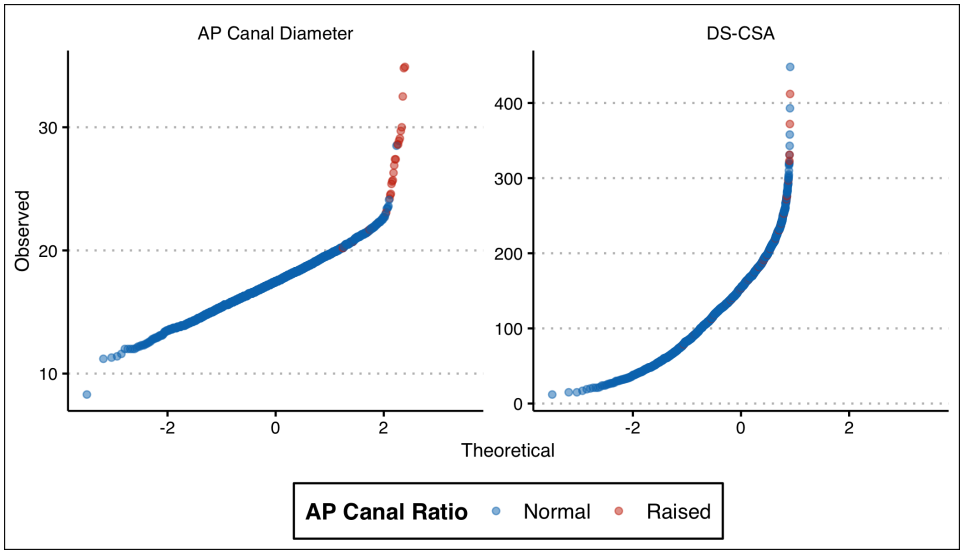


Figure 4.1: Quantile-quantile plots for APD (left) and DS-CSA (right) measurements. Points are coloured according to their AP canal ratio that when raised strongly suggests the presence of isthmic spondylolisthesis. The AP canal ratio is calculated by dividing the measured APD by the APD at L1.

(for a full discussion see Section 1.4.3). The bilateral defect in the neural arch can allow the vertebral body and/or posterior elements to translate relative to each other. In some anterior vertebral body movement predominates, leading to isthmic spondylolisthesis, while in others the posterior elements may sublux dorsally without an associated anterolisthesis, both leading to an widened AP canal diameter at that level. (Ulmer et al. 1995). While pars interarticularis defects are typically difficult to visualise directly on standard lumbar spine MRIs, a widened APD greater than 25% that at L1 (a raised antero-to-posterior canal ratio (APCR)) has been shown to have high diagnostic accuracy for the presence of spondylolysis (Ulmer et al. 1994). Almost all the APD measurements diverging from normal showed a APCR and manual inspection of the images of these confirmed the presence of directly visible pars interarticularis breaks.

The "Schizas" grade

The qualitative A-D grading system proposed by Schizas et al. (2010) is included in Andreisek's core reporting criteria and is an attractive grading system for central canal stenosis as it incorporates both the volume of CSF around the nerve roots of the cauda equina and the presence of effacement of epidural fat — criteria judged of importance in the diagnosis of anatomical LSS by practising clinicians (Mamisch et al. 2012). The Schizas grade has been shown to be more resistant to distortion of grade due to changing axial slice angulation when compared to the DS-CSA (Henderson et al. 2012).

Grading is performed on axial T2 images using the following definitions: grade A — clear CSF visible around the cauda equina nerve roots, grade B — individual nerve roots are visible but fill the entire dural sac and the remaining visible CSF signal is grainy, grade C — individual nerve roots are no longer visible and there is no observable CSF signal within the dural sac, and grade D — as per grade C but with complete effacement of the posterior epidural fat. Example images are provided in figure 4.2. Grading was performed on slices selected by the same methodology as described for the DS-CSA.

The inter-rater reliability was good with a κ of 0.77 (mean agreement between observers 83.2%). The intra-rater reliability was excellent with κ of 0.91 (93.6% agreement between observations).

Discussion between observers and manual inspection of levels with markedly discrepant measurements suggested discrepancies often occurred secondary to effect of CSF pulsation artefacts. The CSF within the dural sac is known to flow with velocity which varies with the cardiac cycle and increases at segments with relative reductions of the dural-sac cross-sectional area (Chun et al. 2017). Prominent flow can cause reductions in CSF signal intensity on T2 weighted images by movement of unexcited protons into the imaged slice at the time of signal acquisition. This effect can cause confusion between grade by obscuring the contrast between nerve roots and CSF — see figure 4.3 — and was thought to be most pronounced at levels close to the boundary between A and B grades.

For the purpose of correlation analyses in the text below the Schizas grades were converted to a linear numeric scale (1 to 4) corresponding to the A to D grades. This process also allowed a mean Schizas grade to be found per participant.

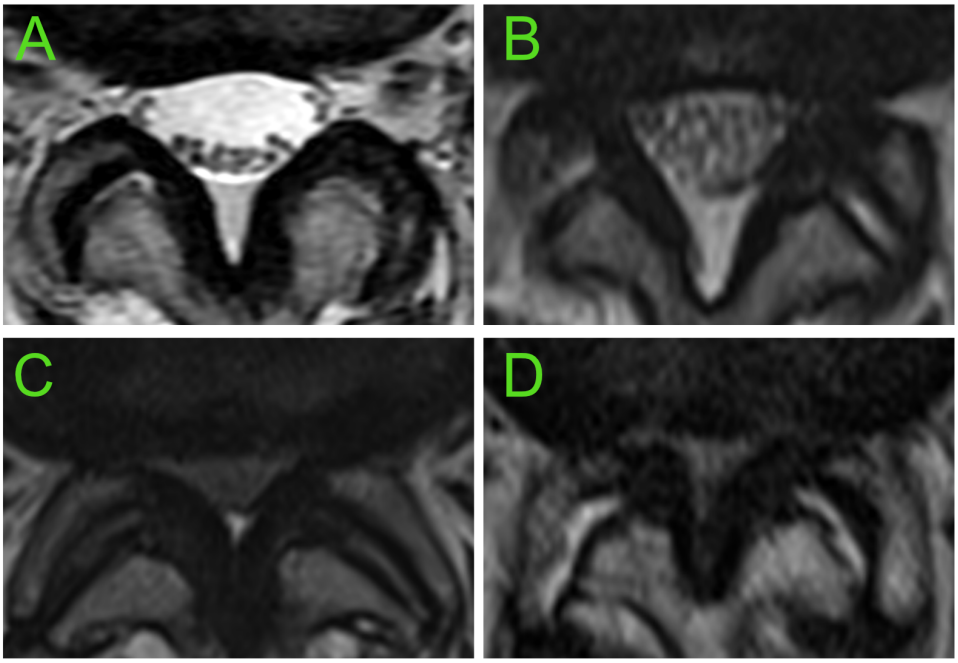


Figure 4.2: Schizas grade example images.

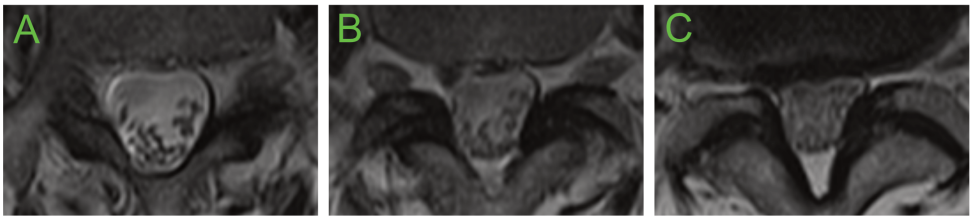


Figure 4.3: Examples of levels presenting difficulties when assessing Schizas grade. All images are 5 mm axial slices from the same patient and vertebral level, moving from the foraminal level on the left (A) inferiorly to the disc level on the right (C). In A the nerve roots are clearly defined with clear CSF signal. Just above the disc (B) the CSF signal has become grainy with poorer nerve root definition. In C the CSF signal has become very grainy with inhomogeneous flow void filling the central aspect of the dural sac simulating the appearance of nerve roots filling the sac completely, but the cross-sectional area of the sac shows minimal change from A to C. Should image C be graded as Schizas A or B?

4.2 Changing canal size with spinal level

There was a clear relationship between the DS-CSA and spinal level, both when looking at all measured levels (Kruskal-Wallis $p < 0.001$) but also when looking purely at levels judged not to show qualitative signs of stenosis (Schizas grade A only, Kruskal-Wallis $p < 0.001$) — figure 4.4.

The median DS-CSA decreased moving inferiorly, before increasing again at L5, consistent with the description of spinal canal size in Bogduk (2012). The median DS-CSA reached its most narrow at L3 in levels without qualitative features of stenosis and at L4 when considering all measured levels (representing L4 as the most common site of stenosis, as explored below).

Similarly, the APD also showed a relationship with spinal level, again both when considering all measured levels (Kruskal-Wallis $p < 0.001$) and for only levels judged as not showing qualitative signs of stenosis (Kruskal-Wallis $p < 0.001$). The smallest median APD was found at L3 (figure 4.4) in both groups.

Anatomical lumbar canal stenosis, as defined using the commonly used threshold DS-CSA measurements of 100 mm^2 and 75 mm^2, was most common at the L4 level and relatively less common at L5 and L1 (table 4.1). These findings were matched by the distribution of Schizas grades, where almost all grades C and D were seen between L2 and L4. In contrast, almost no APD measurements met Verbiest's definitions for relative or absolute stenosis.

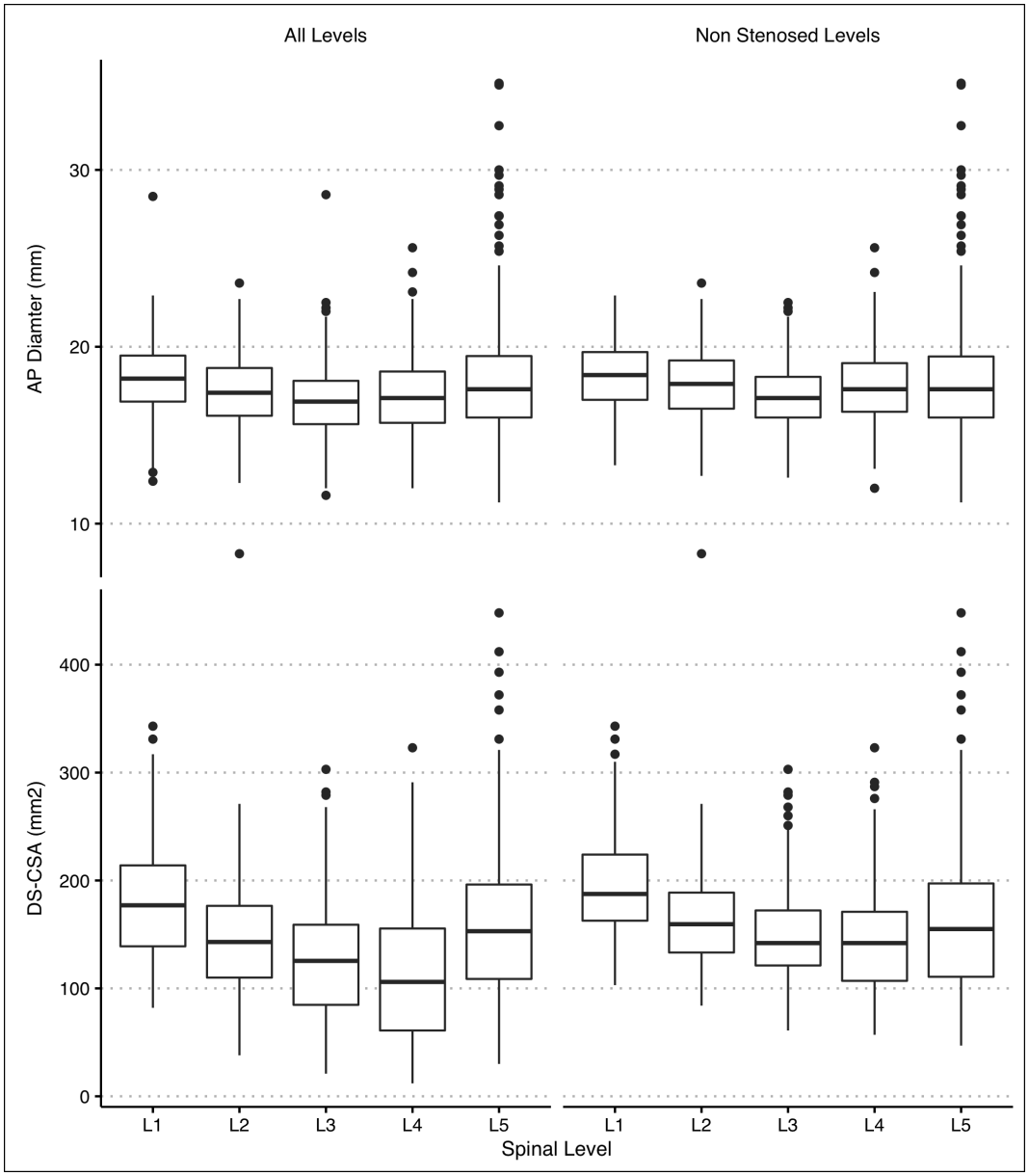


Figure 4.4: Relationship between quantitative measurements of the central spinal canal and spinal level. The plots are divided into those showing data from all performed measurements, and those only from spinal levels judged not to show qualitative signs of central canal stenosis (Schizas grade A).

4.3 Relationship between measurements

There was a moderate correlation between the measured APD and DS-CSA for the same spinal level, with a Pearson's correlation co-efficient of 0.52 (figure 4.5), suggesting the variability in the APD explained around 27% of the variability in the DS-CSA.

There was a strong relationship between Schizas grade and DS-CSA at the same spinal level, with a negative Pearson correlation of 0.61 ($p < 0.001$) between the variables. Similar to the original data presented by Schizas et al. (2010), there was considerable overlap in DS-CSA ranges between grades. A negative trend in median DS-CSA measurements moving from superior to inferior levels was observed at levels graded as B or above (figure 4.6).

In contrast, the relationship between the Schizas grade and the APD was less pronounced — while the Pearson's correlation between these variables was -0.23, this was likely secondary to prominent upper outliers for levels graded as Schizas A — no clear visual trend was present when considering the median or quartile values for the APD between Schizas grades (figure 4.6).

A multivariate regression model, with Schizas grade as the dependent variable and DS-CSA and spinal level as independent variables, had an R^2 of 0.40, suggesting these variation in these variables together account for around 40% of the variability in the Schizas grade. Residual sources of variability in the DS-CSA between grades likely include slice angulation relative to the dural-sac, measurement error and degree of CSF pulsation artefact present on the scan.

Table 4.1: The distribution of central stenosis by spinal level using varying definitions of stenosis. Values provided are percentages of all measurements at that spinal level.

Spinal Level	DS-CSA		APD		Schizas
	< 100 mm ^{2} (%)	< 75 mm^2 (%)	< 12 mm (%)	< 10 mm (%)	$>$ Grade C
L1	3.63	0.00	0.00	0.00	3.11
L2	21.86	10.04	0.24	0.24	15.77
L3	33.74	18.69	0.72	0.00	17.23
L4	46.99	32.77	0.24	0.00	27.23
L5	19.12	9.07	1.21	0.00	1.22

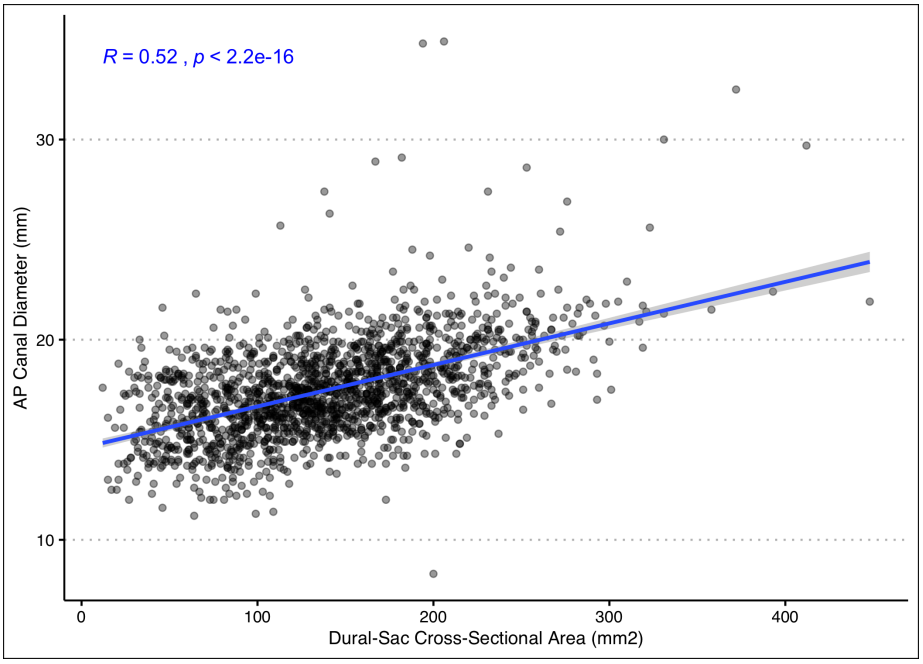


Figure 4.5: The relationship between the APD and DS-CSA measurements at the same spinal level.

4.4 Relationship to demographic variables

The relationships between demographic variables and the measurements of the central canal is illustrated in figures 4.7, 4.8 and 4.9.

A weak negative Pearson's correlation was present between the participants age and their minimum DS-CSA ($R = -0.16$, $p = 0.001$) but there was no corresponding significant correlation with the mean DS-CSA per participant. Interestingly, the correlation between minimum DS-CSA and age was stronger when considering only NC participants ($R = -0.18$, $p < 0.001$) but was insignificant for LBP participants. Age also showed a weak positive correlation with the maximum Schizas grade per participant ($R = 0.18$, $p < 0.001$), with mean age increasing from 73.7 years in those with a maximum Schizas grade of A to 76.9 years in those with a maximum Schizas grade of D. Similarly, the mean Schizas grade was also weakly positively correlated with age ($R = 0.15, p = 0.002$). No significant correlation between age and mean or minimum APD was identified.

An increasing BMI was weakly associated with a reduction in mean DS-CSA ($R = -0.118$, $p = 0.016$) but there was no significant correlation with the minimum DS-CSA. Similarly, BMI weakly correlated with the mean Schizas grade ($R = 0.13$, $p = 0.007$) but not with the maximum Schizas grade. Both mean and minimum APD per participant showed no relationship with participant BMI.

Height was weakly positively correlated with mean DS-CSA ($R = 0.10$, $p = 0.04$), but there was again no significant correlation with the minimum DS-CSA. Height did positively correlate with both mean ($R = 0.13$, $p = 0.01$) and minimum APD ($R = 0.10$, $p = 0.04$). No correlation between the mean or maximum Schizas grade and height was found.

No significant difference in any measurements of the central canal was identified between gender or ethnicity groups.

The mean Schizas grade was significantly smaller in individuals with NC who had a preexisting MRI study compared to the those who had a MRI specifically for the purposes of the trial — 1.38 (confidence interval (CI): 1.32 - 1.45) vs 1.48 (CI: 1.41 — 1.54, Man-Whitney, $p < 0.03$). There was no significant difference in the mean or minimum DS-CSA; mean or minimum APD; or maximum Schizas grade between these participant groups.

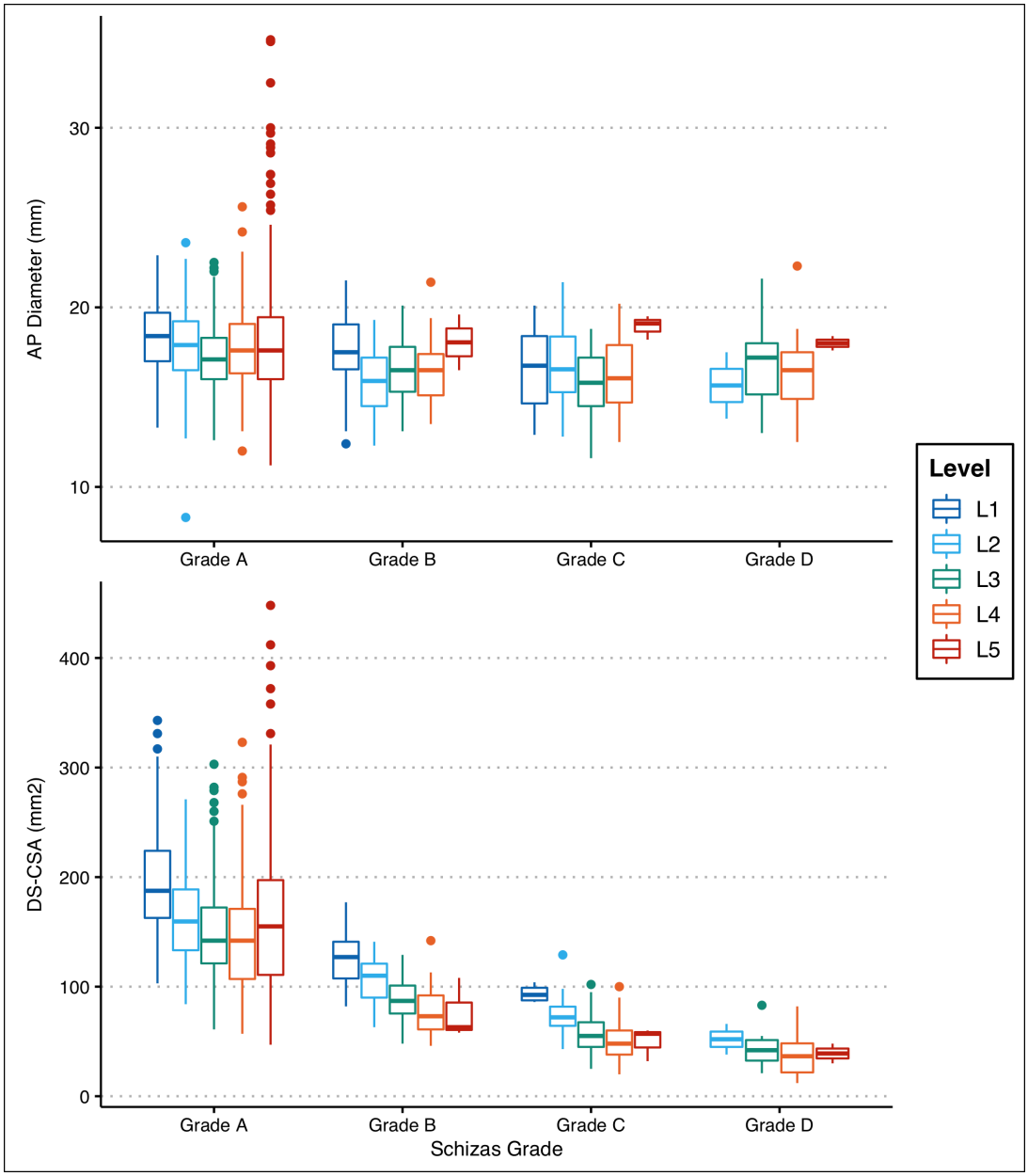


Figure 4.6: The relationship between the Schizas grade and the DS-CSA and APD measurements at the same spinal level.

4.5 Differences between NC and LBP imaging groups

Significant differences between the distributions of both the mean and minimum DS-CSA per participant were identified between NC and LBP groups (table 4.2, figures 4.10). NC participants had an smaller minimum lumbar canal DS-CSA by an average of 36.3 mm^2 and a smaller mean DS-CSA across all measured spinal levels by 29.8 mm^2. Despite this, the degree of overlap in distributions meant the DS-CSA measurements did not clearly distinguish between NC and LBP groups, with AUC values of 0.73 and 0.69 for the minimum and mean DS-CSA respectively. The difference between minimum and mean DS-CSA AUC values was not statistically significant (figure 4.13).

The Youden's J statistic was used to find the optimum diagnostic thresholds (minimising false positives and false negatives) for the both the minimum and mean DS-CSA based on their respective ROC curves (table 4.3). The optimum minimum DS-CSA threshold was calculated as 91.5 mm^2, sitting between the 75 and 100 mm^2 thresholds commonly used in the literature and giving a sensitivity of 63.7% and specificity of 73.0%.

Significantly lower mean and minimum APD measurements per participant were identified in the NC group compared to the LBP group, but the absolute difference in measurements between groups was small with pronounced overlap of the distribution of measurements. The minimum APD per participant was on average 0.7 mm smaller in NC participants compared to LBP groups.

The area under the ROC curve was 0.63 and 0.62 for the mean and minimum APD respectively, again indicating the measurement did not effectively separ-

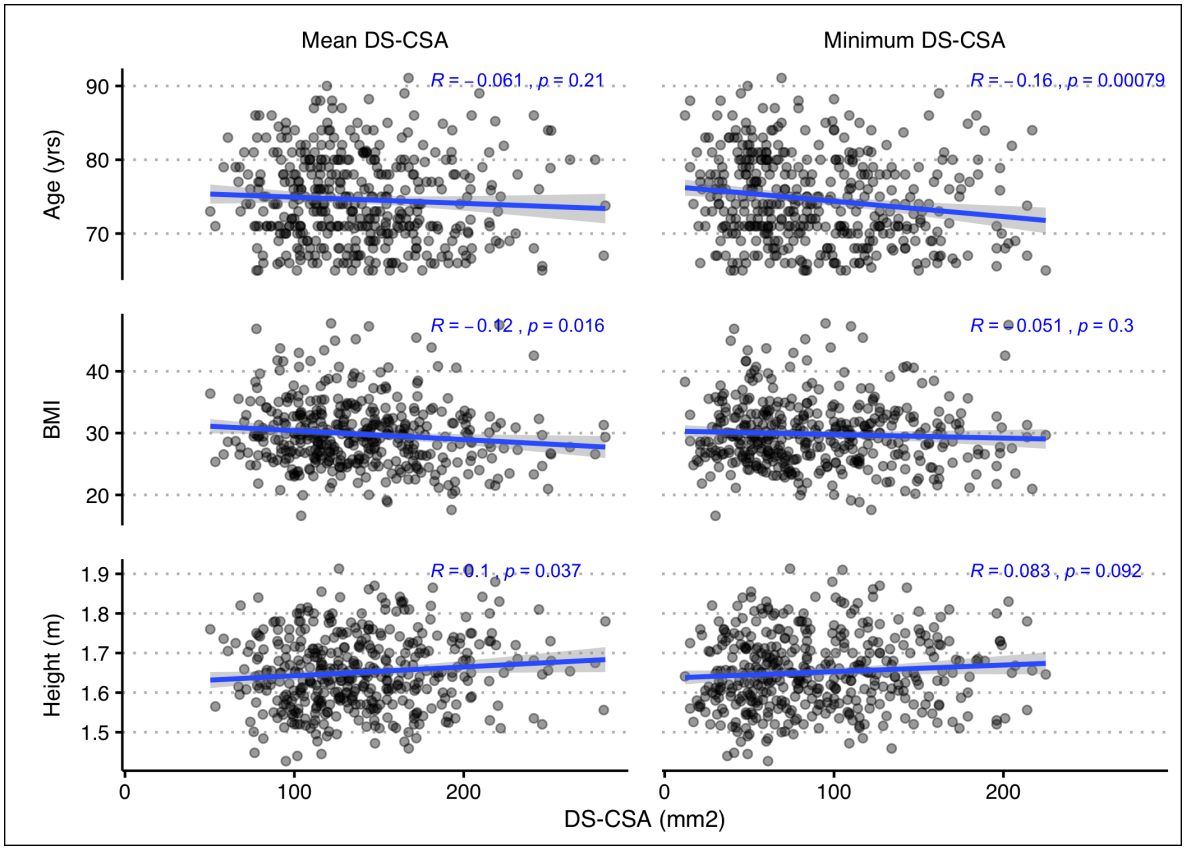


Figure 4.7: Correlation analyses between the DS-CSA and participant age, BMI and height.

Table 4.2: Comparison of mean and minimum DS-CSA and APD and mean Schizas grade (considered as a 1 – 4 point scale) between NC and LBP participants. CI — 95% confidence interval. p values given are Man-Whitney U-test results.

Variable	NC Participant				LBP Participant				p	ROC Analysis			
	Mean	CI			Mean	CI				AUC	CI		
Mean APD	17.4	17.2	–	17.6	18.3	17.8	–	18.7	< 0.001	0.63	0.56	–	0.70
Minimum APD	15.7	15.5	–	15.9	16.4	15.9	–	16.9	< 0.003	0.62	0.54	–	0.69
Mean DS-CSA	134.2	129.9	–	138.5	164.0	152.6	–	175.3	< 0.001	0.69	0.62	–	0.76
Minimum DS-CSA	83.8	79.2	–	88.5	120.1	109.1	–	131.1	< 0.001	0.73	0.67	–	0.79
Mean Schizas grade	1.43	1.39	–	1.48	1.13	1.07	–	1.18	< 0.001	0.71	0.66	–	0.76

ate NC and LBP groups. These AUC values were significantly smaller than the minimum DS-CSA (DeLong, $p = 0.05$ and $p = 0.03$ respectively) but were not significantly smaller than the mean DS-CSA.

The optimum APD threshold for separating the two imaging groups was 17.1 mm, which has an accuracy of 73.6% but a specificity of only 39.7%. As noted above, very few participants met Verbiest's criteria for relative and absolute stenosis and hence their accuracy for separating the groups was poor.

The distribution of maximum Schizas grades was significantly different between the NC and LBP groups (Chi-squared, $p < 0.01$) with more grade C and D scores and fewer Grade A scores in the NC group (figure 4.12). The AUC for the maximum Schizas grade per participant was 0.70 (CI: 0.65 – 0.76), with no significant improvement compared to the mean or minimum DS-CSA AUC. The optimum threshold for the maximum Schizas grades was found to be B or above, which had a sensitivity of 73.0% and specificity of 62.1% for the presence of neurogenic claudication.

The mean Schizas grade was significantly smaller in the NC group (Man-Whitney, $p < 0.001$) with AUC of 0.71. This AUC was not significantly different compared to the maximum Schizas AUC. Interestingly, the ROC curve for the mean Schizas grade showed relative asymmetry favouring specificity (see figure 4.14). Its optimum cutoff (1.45) gave a specificity of 90.5%, with associated sensitivity of 46.4% and accuracy of 53%.

In order to investigate the potential role of multilevel stenosis in neurogenic claudication, the number of spinal levels per participant meeting various definitions of central stenosis were counted and ROC curves were produced (figure 4.15). No assessed threshold for counting stenotic levels had a significantly higher AUC compared to the ROC curves for minimum DS-CSA or maximum Schizas grade.

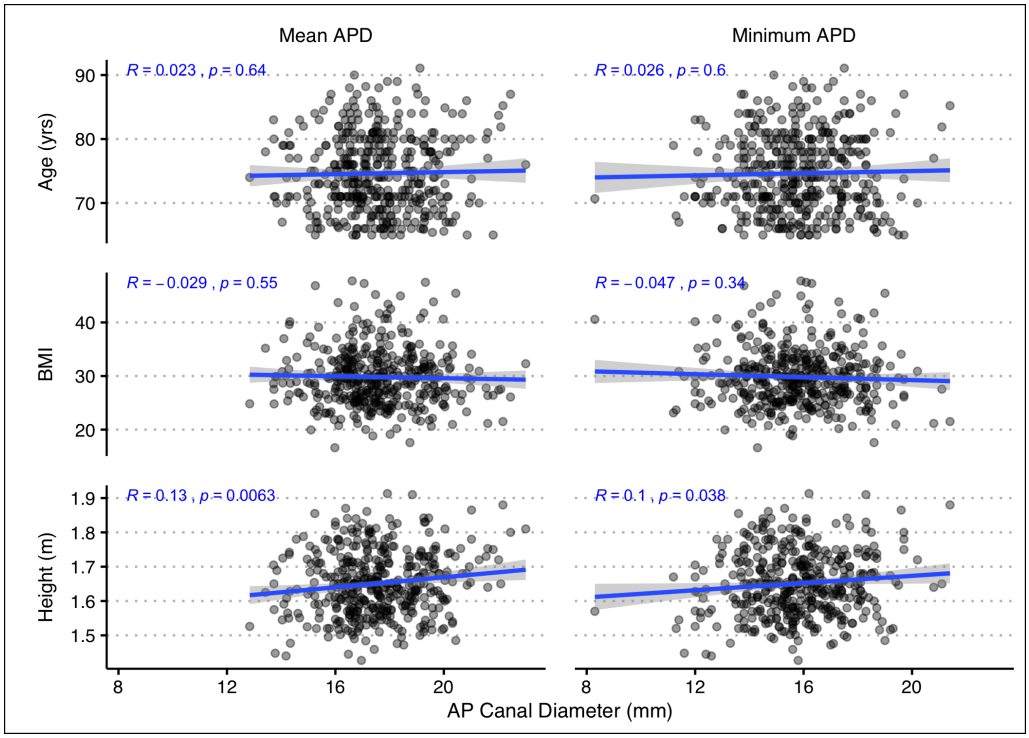


Figure 4.8: Correlation analyses between the APD and participant age, BMI and height.

Table 4.3: Various threshold measurements and their sensitivity and specificity for distinguishing between NC and LBP participant groups. Those measurements marked optimum are the point on the ROC curve with maximum Youden's J statistic.

Measurement	Threshold	Sensitivity (%)	Specificity (%)
Minimum DS-CSA	91.5 mm^2 (Optimum)	63.7	73.0
	75.0 mm^2	52.9	84.1
	100.0 mm^2	66.9	61.9
Mean DS-CSA	135.2 mm^2 (Optimum)	56.9	76.2
Minimum APD	17.1 mm (Optimum)	79.7	39.7
	12.0 mm	2.5	98.4
	10.0 mm	0.0	98.4
Mean APD	18.7 mm (Optimum)	78.2	44.4
Maximum Schizas	Grade B (Optimum)	73.0	62.1
	Grade C	48.4	85.7

4.6 Relationship to symptom severity in BOOST participants

Within the NC group (i.e. BOOST participants) bivariate correlation analyses were performed between the mean and minimum DS-CSA and APD per participant and the clinical severity scores detailed in Section 3.2.7. No statistically significant correlation was identified between any variable pair (figure 4.16).

No significant difference existed in any clinical severity score between participants grouped by their maximum Schizas grade. Similarly, no relationship between the mean Schizas grade per participant and clinical severity was identified.

No significant correlation existed between any clinical severity score and the number of levels of central canal stenosis using any of the previously discussed definitions based on the DS-CSA or Schizas grade.

4.7 Other measurements

The "Lurie" Grade

The grading system proposed by Lurie et al. (2008a) is a formalisation of the normal, mild, moderate or severe stenosis descriptions common to general radiology reports. Lurie suggested guidelines of narrowing of less than one third the normal size as mild, one third to two thirds as moderate, and greater than two

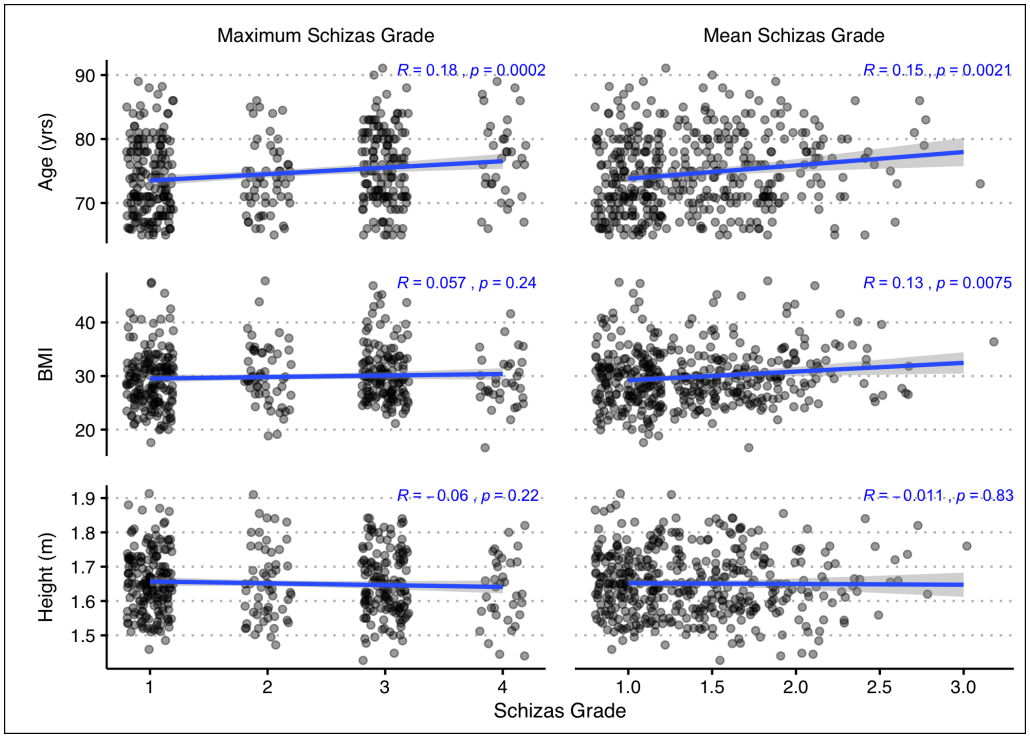


Figure 4.9: Correlation analyses between the mean and maximum Schizas grade per participant (converted to a numeric scale) and participant age, BMI and height. Horizontal jitter has been added to points to reduce overlap.

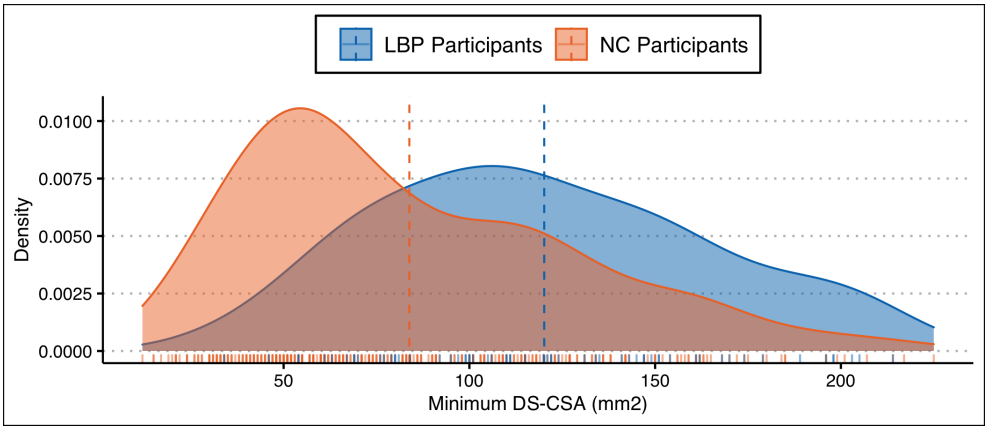


Figure 4.10: DS-CSA measurements. Vertical lines represent mean values.

thirds narrowing as severe stenosis, as judged by the assessor without further definition of normal. The grading system was included in Andreisek's core reporting criteria as a system for measuring both central and neural exit-foramen narrowing. Lurie reported inter-rater κ of 0.73 in his original description of the grading system. In contrast, on the first round of measurement assessment we found an inter-rater κ of 0.29 (68% inter-rater agreement). While observers tended to agree on normal levels, there was wide disparity in grading of levels considered stenotic. Due to this poor performance, no further evaluation of Lurie's grading system for central stenosis was performed.

The sedimentation sign

The sedimentation sign, proposed by Barz et al. (2010), aims to identify potentially clinically significant central canal stenosis by the effect of the stenosis on the distribution of cauda equina nerve roots above and below the level of stenosis. In the normal canal the nerve roots sediment dependently, whereas presence of a stenotic level may prevent this by fixing the entrapped nerve roots within the central aspect of the canal. The sedimentation sign is considered positive if on any axial slice at a spinal level above or below the stenotic level any nerve root (other than the transiting nerve root pair) is within the non-dependent half of the dural sac. Barz reported a inter-observer κ of 0.93, in contrast we found an inter-rater κ of 0.49 in the first round of measurement assessment (74.7% inter-rater agreement). Discussion between observers suggested difficulty in judgement of which slices to include and consistent decisions with regards to nerve roots close to the boundary. No further assessment of the sedimentation sign was performed.

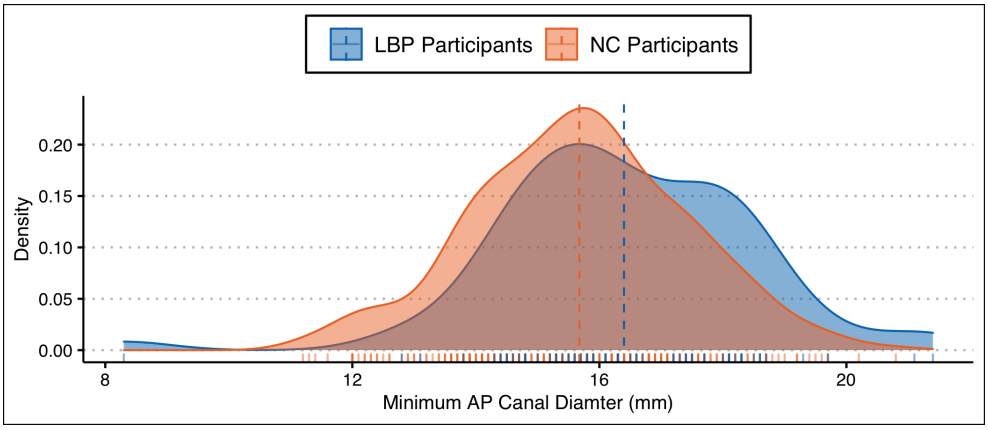


Figure 4.11: APD measurements. Vertical lines represent mean values.

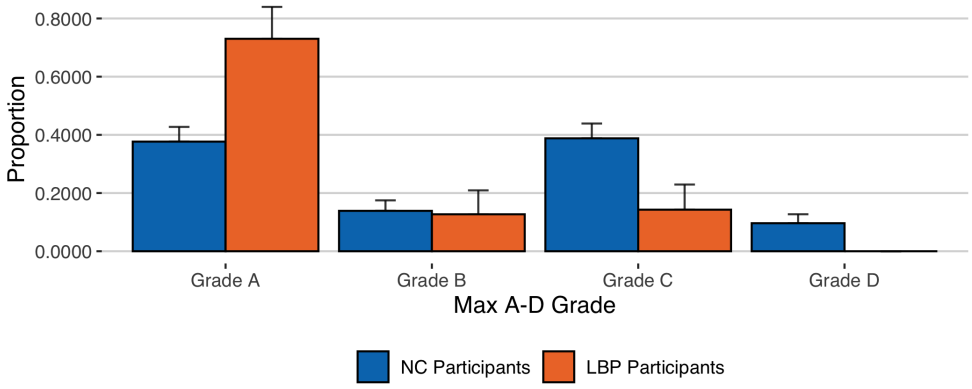


Figure 4.12: Schizas grades

4.8 Chapter summary

Three measurements of central canal narrowing were assessed, the dural-sac cross-sectional area (DS-CSA), antero-to-posterior canal diameter (APD) and the Shizas grade; the former two quantitative measurements of different parts of canal anatomy and the latter a qualitative grading system assessing the degree of entrapment of the cauda equina nerve roots. All three showed good reliability.

The size of the spinal canal followed that described by Bogduk (2012), progressively reducing moving inferior to a minimum at L4/L5 before increasing at L5/S1. Central canal stenosis, as defined using threshold levels of narrowing commonly used in the literature, was most common at L4/5 and seen rarely at L1/2 and L5/S1. As expected, the APD and DS-CSA and Schizas grade were closely related but did not fully explain each others full variability.

There was a negative relationship between NC participant's age and minimum canal size at the disc level as measured by the DS-CSA and Schizas grade, suggesting that canal narrowing progresses with age. The fact that mean DS-CSA did not vary with age raises the possibility that progressive stenosis may only occur at levels at some threshold level of degeneration, consistent with the previously explored hypothesis that a single event (perhaps minor trauma) begins a cascade of degeneration that may affect one level and spare others (see Section 1.4.3). Also as expected, the APD was invariant with age but positively correlated with height, consistent with its suggested role as a marker of developmental canal growth. Increasing BMI was weakly associated with an average reduction in canal diameter at the disc level but not related to the minimum DS-CSA per spine.

The canal was smaller in terms of all three measurements, both at the minimum size per spine and on average across the whole lumbar spine, in neurogenic claudication participants compared to the LBP control group. Despite this, the measurements showed only moderate ability to separate the two participant groups due to the substantial overlap between measurement distributions. The number of stenotic levels also had moderate discriminative ability and there was no relationship between central canal narrowing and any assessed metric of clinical severity in NC participants.

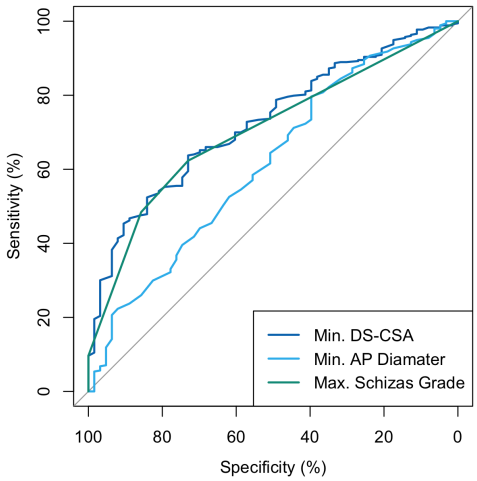


Figure 4.13: ROC analysis investigating the ability of the measurement of the most stenotic level per spine to predict NC or LBP group membership.

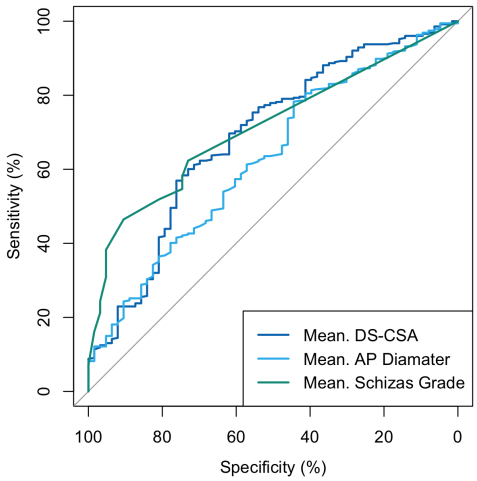


Figure 4.14: ROC analysis investigating the ability of mean DS-CSA, APD and Schizas measurements per spine to predict NC or LBP group membership.

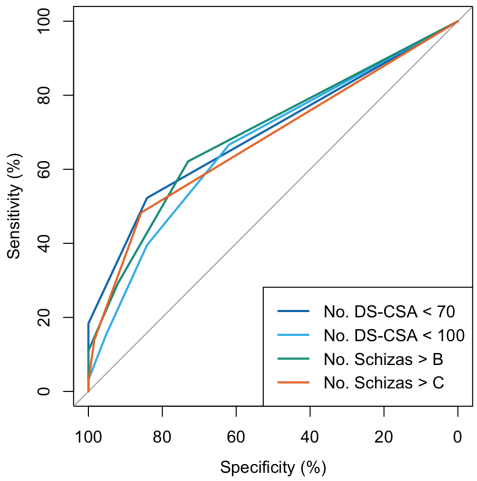


Figure 4.15: ROC analysis investigating the ability of the number of stenotic levels per spine (using varying definitions of stenosis) to predict NC or LBP group membership.

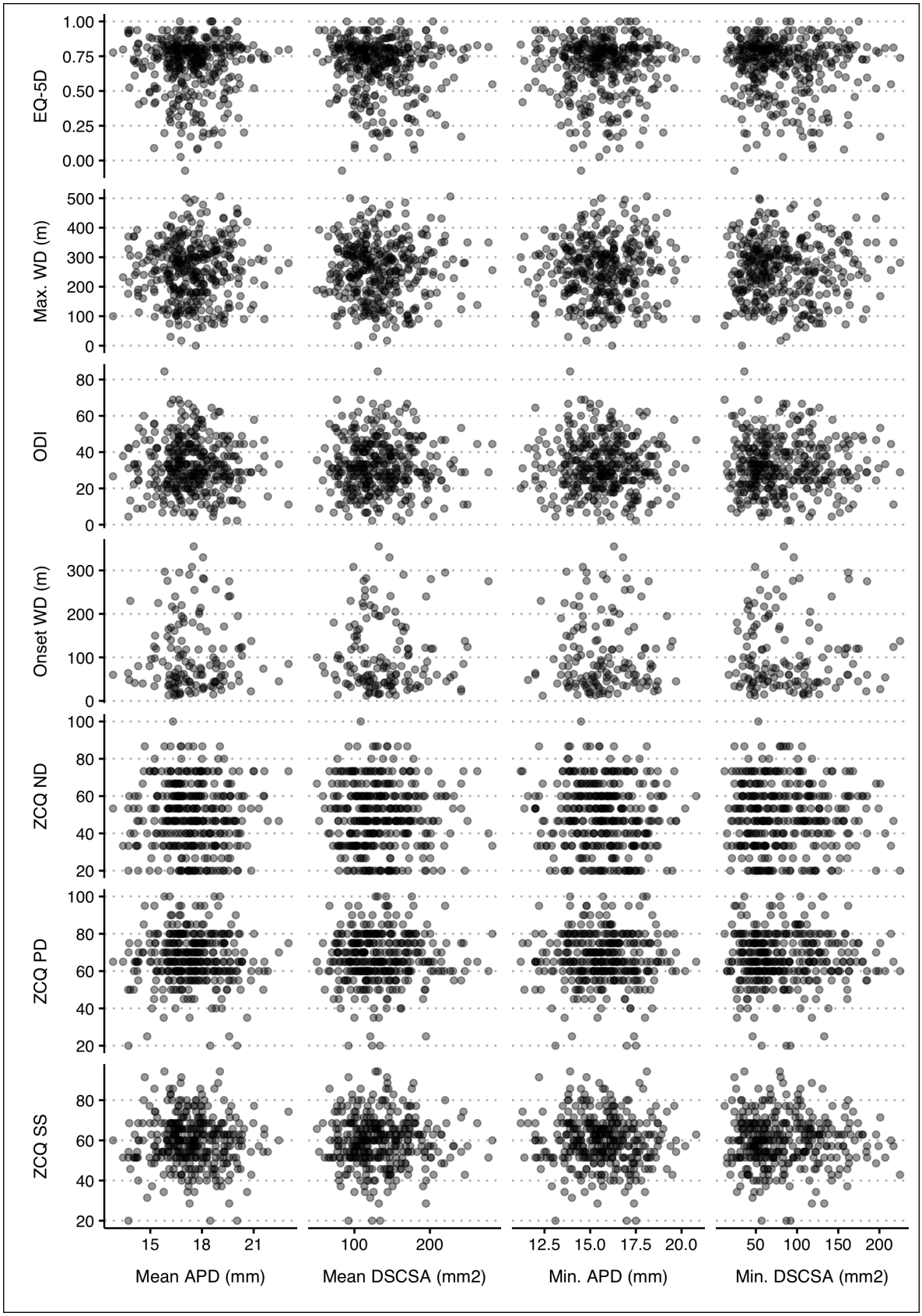


Figure 4.16: Scatter plots showing relationships between clinical severity scores and the mean and minimum DS-CSA and APD per participant. Onset WD refers to the walking distance to symptom onset. No significant correlation was identified between any variable pair.

Chapter 5

The Lateral Recess

5.1 Measurement method and reliability

Lateral recess depth (LRD)

The LRD was a measurement pre-specified for use in the BOOST MRI based sub-group analysis of treatment effect and was included as a research measurement in Andreisek's core reporting criteria.

As discussed in § 1.3.3, the LRD was originally described as a measurement taken at the pedicular level across the outer aspect of the bony lateral recess — from the medial aspect of the superior articular facet anteriorly in the sagittal plane to the posterior vertebral body cortex (figure 1.4, measurement 8). For the purposes of this thesis it was decided to alter this measurement to be measured at the disc level with the following technique: a digital ruler was used to measure between the facet joint capsule/ligamentum flavum immediately anterior to the most medial aspect of the superior articular facet anteriorly in the sagittal plane to the posterior disc annulus. The measurement was made on axial T2 images at the disc level, on the slice with the most significant degree of dural-sac narrowing. Given its anatomical relations, this disc-level LRD would be expected to decrease with both increasing apophyseal joint/ligamentum flavum hypertrophy and increasing disc bulge/protrusion and so more closely relate to the risk of transiting nerve-root entrapment within the ligamentous portion of the lateral recess.

A similar disc-level version of the LRD was described by Chung et al. (2000) — the sub-articular sagittal diameter of the canal (SSDC), discussed further in § 5.7. The disc-level LRD described above differs from the SSDC by being fixed at the medial border of the ligamentous portion of the lateral recess, rather than moving relative to the borders of the lateral recess depending upon the morphology of the superior articular facet (which varies across different spinal levels — figure 5.1).

The LRD was found to have an good inter-rater ICC of 0.65 and, when applied to the full MRI dataset, an intra-rater ICC of 0.87. Discussion between observers suggested the measurement became difficult to use when the medial end of the apophyseal joint line became obscured by osteophytosis and when the degree of hypertrophy or disc bulge made the lateral recess very small. In the latter case observers tended to measure the LRD as 0 mm, creating a relative paucity of very small measurements not equal to 0. This effect can be seen in the various graphs demonstrating the distribution of LRD measurements below.

The LRD was not normally distributed (Shapiro-Wilk, $W = 0.97$, $p < 0.001$), with sharp deviation from normal at the lower end of the distribution where measurements became 0 mm (figure 5.2).

The "Bartynski" grade

The qualitative grading system proposed by Bartynski et al. (2003) aims to assess for entrapment of the transiting nerve roots within the lateral recess and is included in Andreisek's core reporting criteria. The grading system defines 4 grades and is judged on axial T2 weighted images at the mid-disc level: 0 — normal lateral recess; 1 — reduced size of the lateral recess with trefoil shape; 2 — reduced size of the lateral recess with the transiting nerve root judged compressed, but without complete obliteration of the recess; 3 — no CSF or space visualised in the lateral recess and compression of the transiting nerve root. The original description also includes a statement that for grades 2 and 3 the nerve root may be displaced medially of its normal position, though it is unclear if this is in addition to, or in place of, any nerve root compression. Examples of different grades are provided in figure 5.3

During initial testing of the grading system the inter-rater κ was found to be 0.61 (58.0% agreement). Discussion between observers suggested difficulties regarding interpretation of the role of nerve root displacement. For instance, in many cases the lateral recess as defined by the position of the superior articular facet, would appear completely effaced but the transiting nerve root could clearly be identified running medially within the canal without compression (figure 5.4), not clearly fitting into grades 1 or 3. Further difficulties were noted in the subjective nature of the boundary between grade 0 and grade 1.

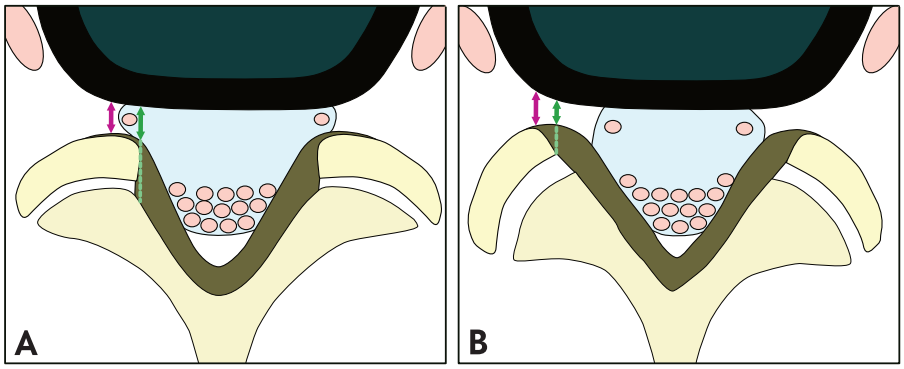


Figure 5.1: Demonstration of disc-level LRD and SSDC measurement techniques both at spinal levels with the more coronally orientated articular facets typical of lower lumbar levels (panel A) compared to the more sagittal facet orientation more typical at upper lumbar levels (panel B). The LRD is shown in dark green and is measured anterior to the medial most aspect of the superior articular facet (marked by the green dotted line). The SSDC is measured anteriorly from the ligamentous covering of the most anterior aspect of the superior articular facet, typically lying more lateral to the LRD measurement. Note that in panel B the lateral aspect of the dural sac lies medial to both measurements as the canal and the dural sac has a more ovid, rather than trefoil cross-section. In fact, some authors consider there to be no true lateral recess in the upper lumbar spine where the canal has this morphology (Wilmink 2010b).

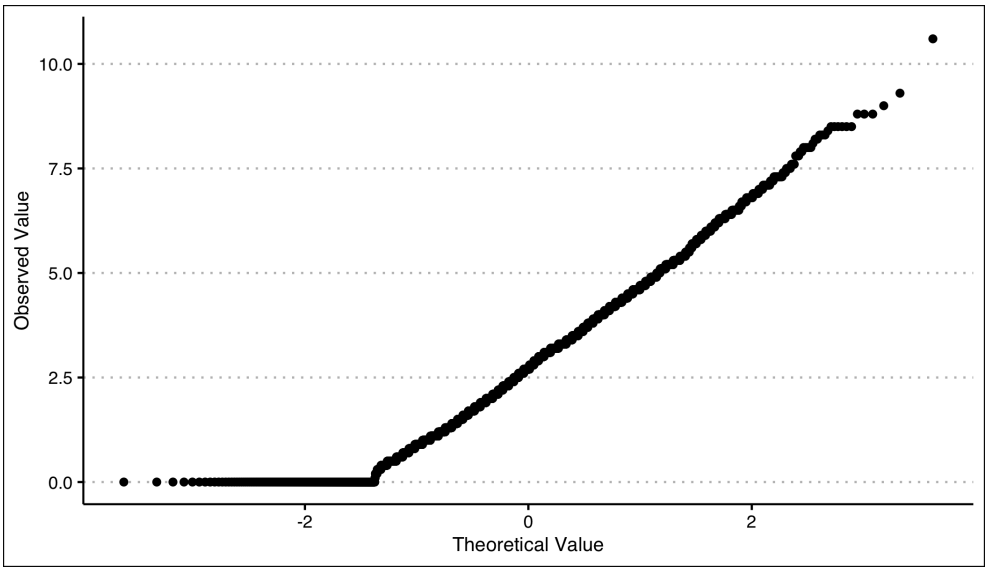


Figure 5.2: Quantile-Quantile plot of LRD measurements.

In a second round of testing the grading system was modified to exclude the statement about nerve root displacement and the intra-rater κ increased to 0.70 (71% agreement). On application to the full dataset this method gave an intra-rater κ of 0.82 (73 % agreement).

For the purposes of correlation analyses in the text bellow the Bartynski grades were converted to a 4 point linear numeric scale.

Disc herniation

Measuring the presence and effect of disc herniations proved difficult. Initial attempts to capture this information used a multi-point grading system which measured both the presence, site and degree of nerve root compression if present at each level. Unfortunately inter-rater reliability scores for this system were very poor, and its grading added substantially to the amount of time required for each study.

The initial grading system was replaced by a single question relating to whether a protrusion/extrusion was both present, as defined by the lumbar disc nomenclature (version 2, Fardon et al. 2014), and was causing compression or displacement of a cauda equina nerve root. The question was answered for each level at which axial images were available. The inter-observer κ score was poor (0.28) but with good percentage agreement (85.6%). In contrast the intra-observer reliability was good with κ of 0.71 (96.2% agreement).

Manual inspection of discrepancies suggested that while observers generally agreed about normal levels, smaller/borderline disc protrusions were often picked up by only 1 or 2 observers, disproportionately affecting the κ score.

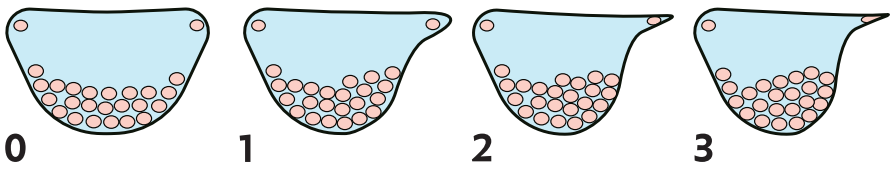


Figure 5.3: Illustration of the Bartynski grading scheme of lateral recess nerve root entrapment. Each diagram shows a dural-sac in cross-section with its associated nerve roots. From 0 to 3 there is progressive stenosis of the left lateral recess. 0 — the lateral recess is normal; 1 — the left lateral recesses has taken on a trefoil shape but there is no indication that the left transiting nerve root is compressed. 2 — there is loss of CSF space around the left transiting nerve root which has become flattened due to the compression. CSF space remains visible lateral to the nerve root; 3 — the left lateral recess is obliterated with the transiting nerve root flattened and with no clear space lateral to it.

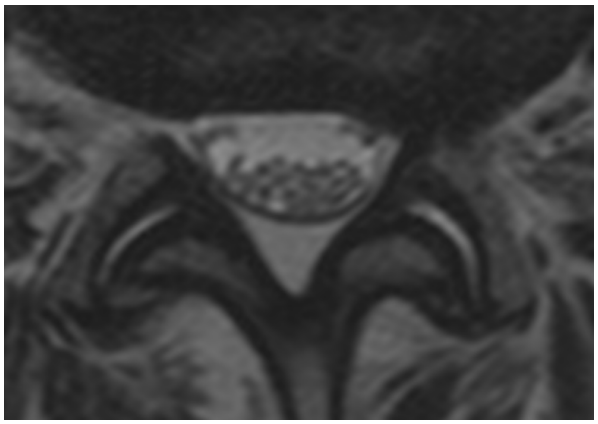


Figure 5.4: Both left and right nerve roots are medially placed relative to the superior articular facet without evidence of compression. Despite this left lateral recess depth is 0 mm.

5.2 Changing lateral recess size with spinal level

The LRD had a relationship with spinal level, both when considering all measured levels (Kruskal-Wallis, $p < 0.001$) and those without qualitative features of stenosis (Bartynski grade 0, Kruskal-Wallis, $p < 0.001$). The median LRD progressively decreased between L1 and L4 before increasing again at L5 (see figure 5.5). This is consistent with the increasing antero-lateral displacement of the apophyseal joints moving inferiorly through the lumbar spine (Bogduk 2012).

A lateral recess depth of 3 mm is commonly used in the literature as a threshold below which the lateral recess is considered stenotic (Steurer et al. 2011). Using this definition, 55.6% of all measured lateral recesses showed anatomical stenosis. Similarly 53% of all measured levels were judged qualitatively to be abnormally narrowed (Bartynski grade of 1 or greater) with 25% showing qualitative features of transiting nerve root compression (Bartynski grade 2 or greater).

Lateral recess stenosis, by both the qualitative and quantitative definitions above, was most common at L4 and L3, and relatively rarer at L1 (see table 5.1). Recesses with qualitative evidence of transiting nerve root compression were similarly distributed but showed a relatively lower incidence at L5 than would be expected given the incidence of recesses judged stenotic at that level according to the 3 mm LRD threshold. This latter finding may well represent the relatively lateral positioning of the superior articular facet at this level compared to the dural sac and transiting nerve root.

Overall 32.9% of all MRIs showed a significant disc herniation (a protrusion or extrusion that causes deviation or compression of a nerve root) affecting at least one visualised level. Significant protrusions were most common at the L5 and L4 levels (where 19.7% and 10.6% of visualised levels showed significant protrusions respectively).

5.3 Relationship between measurements

There was a medium negative correlation between the Bartynski grade and the lateral recess depth: $R = -0.64$, implying the LRD explains 41% of the variability in the grading ($p < 0.001$, figure 5.6).

Across all assessed lateral recesses there were 18 LRD measurements of 0 mm where the Bartynski grade had been assessed as 0 (i.e. no lateral recess narrowing). Manual inspection of the MRI images for the spinal levels involved suggested that this data largely represented grading errors — these levels had

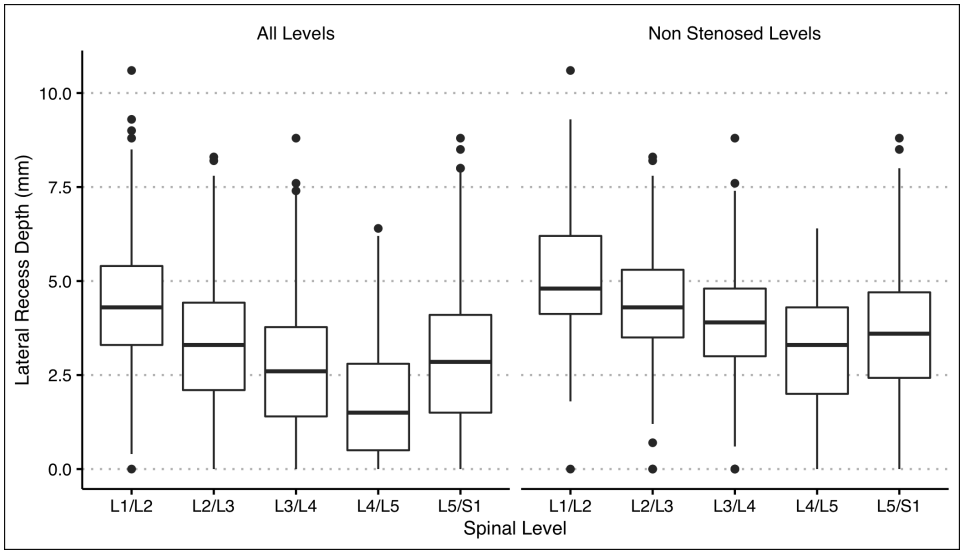


Figure 5.5: The relationship between LRD and the spinal level, both for all measured levels and for those not showing qualitative features of stenosis (Bartynski grade 0).

Table 5.1: Distribution of lateral recess stenosis by spinal level set using varying definitions of stenosis.

Spinal Level	LRD $<$ 3 mm (%)	Bartynski Grade $>$ 0 (%)	Bartynski Grade $>$ 1 (%)
L1	18.1	26.4	7.4
L2	43.0	49.6	19.3
L3	61.1	61.3	26.3
L4	77.9	73.9	42.9
L5	53.8	42.2	17.9

a dural sac whose lateral margin lied medial to the bony/ligamentous boundaries of the lateral recess avoiding compression despite the lateral recess narrowing, and should have been graded Bartynski 1. This ability of the transiting nerve root and/or dural sac to lie outside the lateral recess also explains the presence of 0 mm LRD measurements in the Bartynski Grade 1 category.

A ROC analysis was performed to assess the ability of the LRD to predict qualitative features of nerve root compression (figure 5.7). The commonly used 3 mm LRD definition of lateral recess stenosis was associated with a 95.9% sensitivity and 57.8% specificity for qualitative features of nerve root compression (i.e. Bartynski Grades 2 and 3). In contrast Youden's J statistic suggested an optimal LRD threshold of 2.0 mm, (minimising false positives and false negatives) with sensitivity of 81.8% and specificity of 77.4%.

Perhaps unsurprisingly, given the same degenerative processes contribute to both lateral and central canal narrowing, the LRD was closely related to measures of central canal stenosis at the same level. The LRD was strongly positively correlated with DS-CSA in a manner dependent upon the spinal level: moving inferiorly there was a decrease in both strength of correlation and the gradient of the line of best fit (0.25 at L1/2 to 0.16 at L5/S1) — figure 5.8. The relationship between the LRD and APD was weaker than that found for the DS-CSA and less effected by the spinal level (figure 5.9), again consistent with the APD being relatively unaffected by age related degeneration at the azygous joint/disk level.

The Bartynski grade was moderately positively correlated with the Schizas grade ($R = 0.68$, $p < 0.001$, figure 5.10) and there was an odds ratio of 63.9 (95% CI 45.6 – 89.4, $p < 0.001$) between the probabilities of central nerve root entrapment (i.e. a schizas grade > B) and transiting nerve root impingement at the same level (i.e. a Bartynski grade > 1).

As expected, the presence of a significant disc herniation was associated with smaller lateral recess diameters (Man-Whitney U, $p < 0.001$), DS-CSAs (Man-Whitney U, $p < 0.001$) and with higher Bartynski grades (Chi-squared $p < 0.001$). Herniations did not have any significant affect on the Schizas grade (suggesting that the majority were not large enough to significantly compress the central dural-sac) and, as expected, did not have a relationship with the APD — table 5.2.

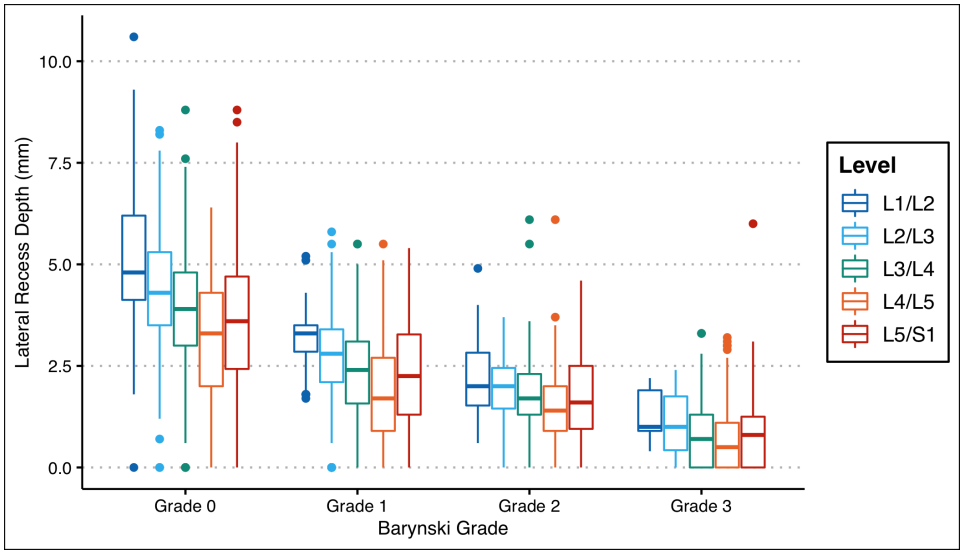


Figure 5.6: The relationship between lateral depth measurements and Bartynski grades of nerve root entrapment in the lateral recess.

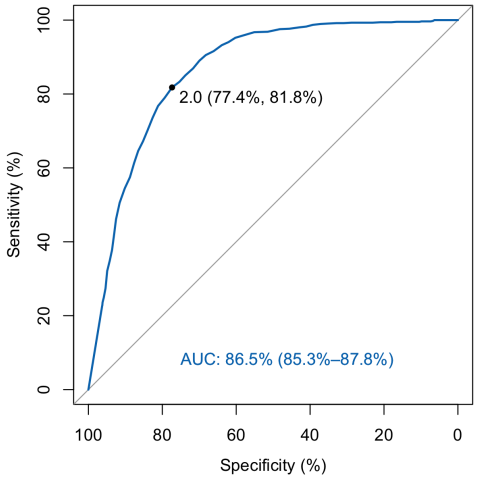


Figure 5.7: ROC analysis — the ability of the LRD to predict a Bartynski grade of 2 or above.

5.4 Relationship to demographic variables

The relationships between demographic variables and the LRD and Bartynski grade are demonstrated in figures 5.11 and 5.12 respectively.

There was a weak negative correlation between the participant age and minimum LRD per participant ($R = -0.15$, $p = 0.002$) and weak positive correlations with both the maximum and mean Bartynski grades per participant ($R = 0.16$ and $R = 0.13$ respectively, $p < 0.01$). Similar to the findings seen with the DS-CSA the relationship between the age and minimum LRD was stronger in patients with NC ($R = -0.18$, $p < 0.001$) but insignificant for control participants. No significant correlation was identified between participant age and mean lateral depth.

Participant height had a weak negative correlation with both mean and minimum LRD ($R = -0.22$ and $R = -0.12$ respectively, $p < 0.02$). No significant correlation was identified between participant height and the mean or maximum Bartynski grade.

No significant correlation between participant BMI and either the mean or minimum LRD or mean or maximum Bartynski grade was identified.

Both the minimum and mean lateral recess depth per participant means were smaller in men compared to women (see table 5.3, $p < 0.001$ for both). Similarly the mean of the mean Bartynski grade per participant was larger in men compared to women ($p = 0.018$). In contrast, no significant difference in the distribution of maximum Bartynski grades were seen between gender groups.

No significant difference in any lateral recess measurement was seen between ethnicity groups.

There was no significant difference in age, BMI or height between participants with and without a significant disc herniation and no correlation between age, BMI and height and the number of disc herniations per participant. There was no significant difference in gender or ethnicity distribution between

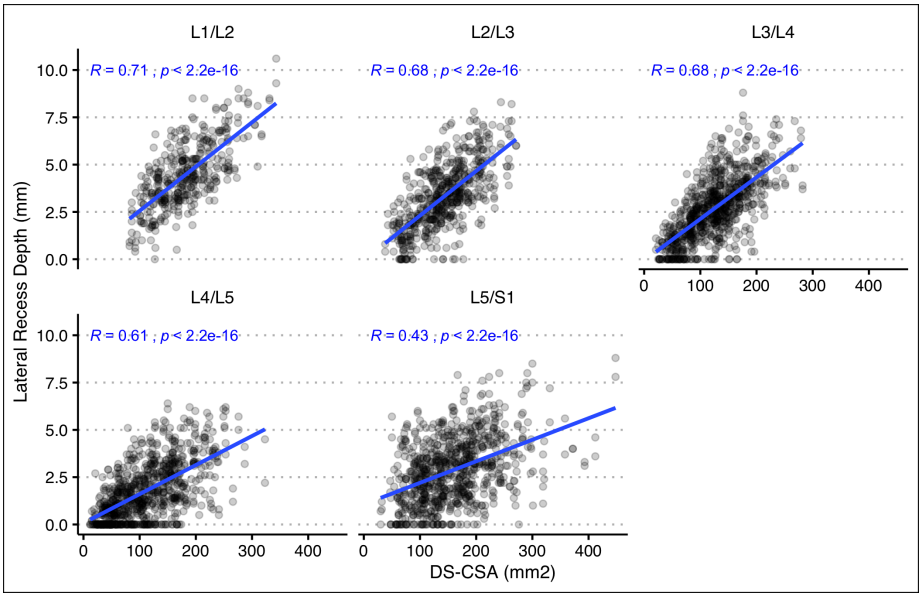


Figure 5.8: Scatter plots of lateral recess depth compared to the DS-CSA.

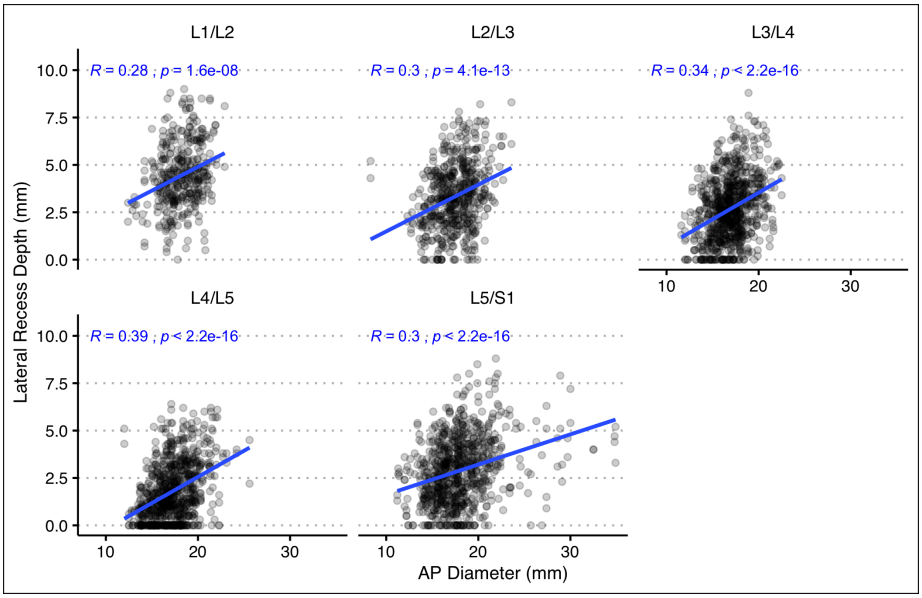


Figure 5.9: Scatter plots of lateral recess depth compared to the APD.

those with and without a significant disc herniation and no significant difference in the distribution of the number of herniations per participant between gender and ethnicity groups.

Among participants with NC (recruited as part of the BOOST trial), the distribution of max Bartynski grades per participant differed between those with a preexisting clinical MRI scan compared to those undergoing a MRI specifically for the purposes of the trial (Chi-Squared, $p = 0.02$) with a higher percentage of the preexisting clinical scan group having Grade 3 maximum scores (65.0% vs 50%). There were no significant differences in the LRD mean or minimum per participant, the mean Bartynski per participant, or the presence of significant disc herniations, between those NC participants with and without a preexisting clinical scan.

5.5 Differences between NC and LBP imaging groups

Both the minimum and mean lateral recess depth per participant was smaller in the NC participants compared to the LBP participants (see table 5.4 and figures 5.13 and 5.14). There was however substantial overlap in the distributions of the measurements comparing the two participant groups and this resulted in a poor ability to distinguish NC from LBP participants based on these measurements, with AUCs of 0.66 and 0.63 for the minimum and mean LRD respectively (figure 5.17). The difference in AUC values

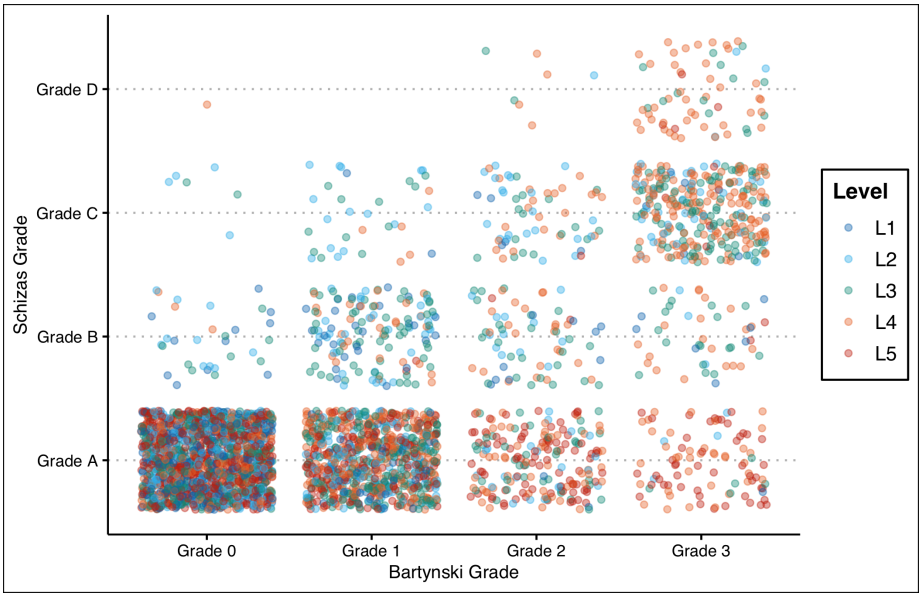


Figure 5.10: The relationship between Bartynski grade and Schizas grade at the same spinal level.

Table 5.2: The effect of the presence of significant disc herniations on measurements of the central canal and lateral recess

Variables	Disc Herniation Present				Disc Herniation Absent				P Value
	Mean	CI			Mean	CI			
LRD (mm)	2.00	1.91	–	2.08	2.90	2.84	–	2.96	< 0.001
DS-CSA (mm^2)	120.49	117.59	–	123.39	141.33	139.34	–	143.32	< 0.001
APD (mm)	17.26	17.13	–	17.38	17.50	17.42	–	17.58	0.074

for the mean and minimum LRD curves was not statistically significant.

Higher mean and maximum Bartynski grades per participant were found in the NC participants. The mode for maximum Bartynski grade per participant was 3 in NC participants, compared to 1 in LBP participants (see figure 5.16, Chi-Squared $p < 0.001$). Again, due to distribution overlap, the Bartynski grade had only a moderate ability to separate NC from LBP participants with AUCs of 0.77 and 0.74 for the mean and maximum grade per participant respectively (figure 5.17). The AUC values for the mean and maximum Bartynski grade did not differ significantly but were both significantly higher than the AUC values for the minimum and mean LRD measurements (Dunn's test, $p <= 0.01$ for all comparisons).

The performance of different diagnostic thresholds of the lateral recess measurements are demonstrated in table 5.5. The optimum threshold for each measurement was found by Youden's J statistic and was 0.6 mm for the minimum LRD and Grade 3 for the Bartynski grading system. The optimum thresholds generally had moderate specificity's but poor sensitivities. In contrast, the 3 mm LRD threshold commonly used in the literature to define lateral recess stenosis had a high sensitivity for NC symptoms but very poor specificity. Similarly, the presence of qualitative features of nerve root compression (Bartynski grade 2 or above), had better sensitivity but poorer specificity compared to the optimum threshold of Grade 3 (complete obliteration of the lateral recess).

To investigate the role of multi-level lateral recess stenosis, the ability of the number of lateral recesses meeting various thresholds of lateral recess narrowing per spine were assessed for their ability to distinguish NC participants from LBP participants (figure 5.18). There was no significant improvement in AUC between the number of narrowed lateral recesses by any of the investigated LRD thresholds and

Table 5.3: Comparison of mean and minimum LRD and mean Bartynski grades between genders.

Variables	Women				Men				p Value
	Mean	CI			Mean	CI			
Min. LRD (mm)	1.3	1.1	–	1.4	0.8	0.7	–	1.0	< 0.001
Mean LRD (mm)	3.1	3.0	–	3.3	2.3	2.2	–	2.5	< 0.001
Mean Bartynski Grade	0.87	0.79	–	0.94	1.04	0.94	–	0.14	0.018

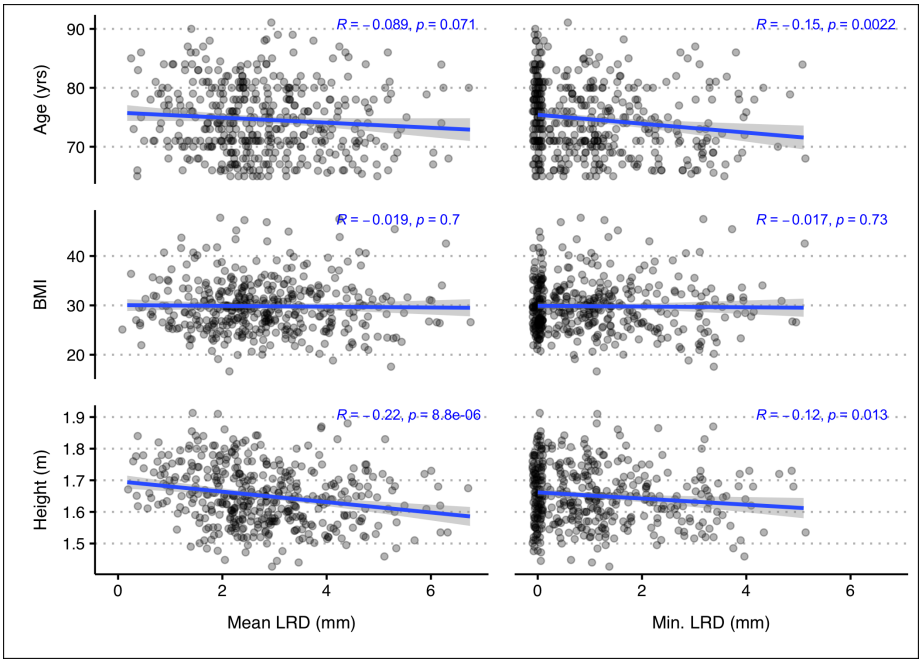


Figure 5.11: The relationship between mean and minimum LRD per participant and participant Age, BMI and height measurements.

Table 5.4: Comparison of lateral recess measurements between NC and LBP participant groups.

Variables	NC Participants				LBP Participants				p Value	ROC Analysis			
	Mean	CI			Mean	CI				AUC	CI		
Min. LRD (mm)	1.0	0.9	–	1.1	1.6	1.3	–	1.9	< 0.001	0.66	0.59	–	0.73
Mean LRD (mm)	2.7	2.6	–	2.8	3.2	2.9	–	3.5	0.001	0.63	0.55	–	0.70
Mean Bartynski Grade	1.03	0.96	–	1.09	0.47	0.37	–	0.56	< 0.001	0.77	0.71	–	0.82

the mean or minimum LRD per spine. Similarly the AUC for the number of lateral recesses showing at least Bartynski Grade 2 or 3 was not greater than the AUC for mean or maximum Bartynski grade per participant.

There was no significant difference in the presence of significant disc herniation or number of significant herniations per spine between NC and LBP participants.

5.6 Relationship to symptom severity in NC participants

No significant correlation was identified between the mean or minimum LRD per participant and any clinical severity variable in NC participants (figure 5.20). Similarly, there was no significant correlation between the mean or maximum Bartynski grade per participant and any clinical severity variable.

There was also no significant correlation between any clinical severity variable and the number of lateral recesses judged stenotic using the various thresholds of lateral recess narrowing explored in the previous section: including LRD < 3 mm, LRD < 0.6 mm, Bartynski Grade $>= 2$, Bartynski Grade $>= 3$. Similarly, there was no significant difference in clinical severity score between multi-level, single-level and no lateral recess stenosis groups divided using the threshold measurements above.

Participants judged as having significant disc herniations appeared to have lower levels of neuro-ischaemic symptoms as judged by the ZCQ neuro-ischaemic domain, leading to an overall lower score for such participants in the overall ZCQ symptom subscale (Man-Whitney, $p = 0.02$ and $p = 0.03$ respectively, table 5.6). No significant difference was observed in the pain domain scores between the groups. Similarly, weak negative correlations were identified between the number of discs demonstrating significant herniations per participant and the ZCQ neuro-ischaemic domain and ZCQ symptom subscale ($R = -0.14$, $p = 0.006$ and $R = -0.12$, $p = 0.02$ respectively, see figure 5.19).

No significant difference in ODI, EQ-5D, walking distance to symptom onset or max walking distance was found between those with and without significant disc herniations. No significant correlation between the number of discs demonstrating herniations per participant and the ZCQ pain domain, ODI, EQ-5D, walking distance to symptom onset and maximum walking distance was identified.

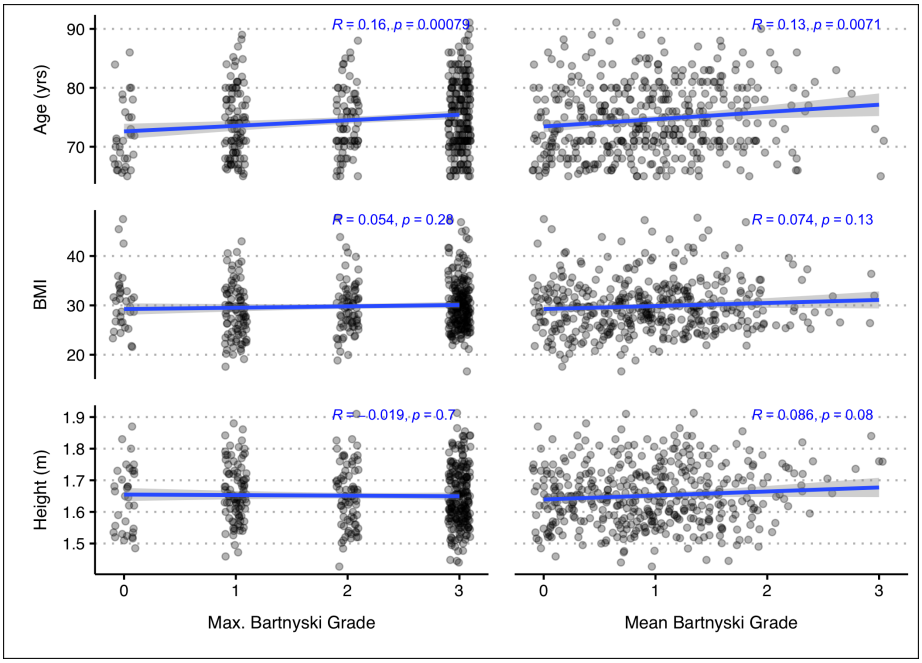


Figure 5.12: The relationship between maximum and mean Bartynski grades per participant (converted to a 4 point numerical scale) and participant Age, BMI and height.

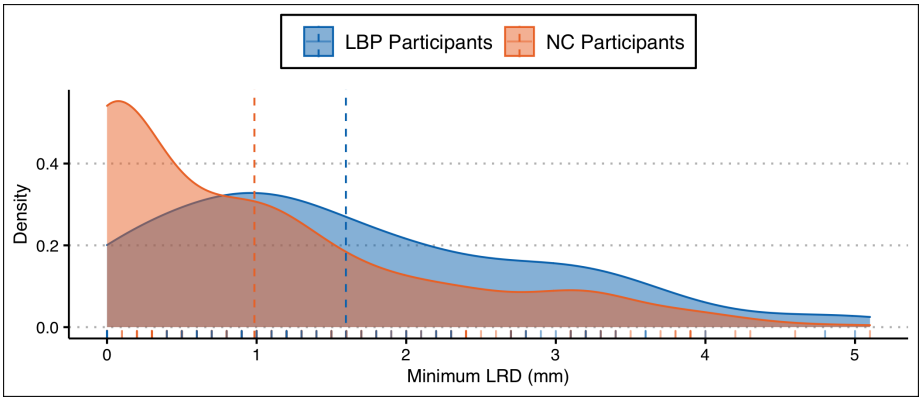


Figure 5.13: Distribution of minimum lateral recess depth measurements per participant. Horizontal lines represent the mean value of the distributions.

5.7 Other measurements

The sub-articular sagittal diameter of the canal (SSDC)

The SSDC was measured from the most anterior aspect of the capsular covering capsule overlying the superior articular facet, anteriorly to the posterior disc annulus (figure 5.1). The SSDC was added as a measurement in the second round of measurement assessment.

We found the SSDC inter-rater to be poor (ICC = 0.31). Measurement discrepancies were most prominent at the L5/S1 disc, where the most anterior aspect of the superior articular facet often lies lateral to the disc giving no clear anterior end to the measurement. Observers had adopted different techniques for dealing with this.

While further standardisation of measurement technique would likely improve the observed ICC, the disc-level LRD had performed well and both measurements assessed similar anatomical aspects of the canal. Hence, no further investigation of the SSDC was undertaken.

Ligamentum flavum thickness

Two measurement techniques were assessed for measurement of the ligamentum flavum thickening, measuring on axial T2 images at mid-disc level at the level of the azygous joint space, and measuring at the mid point of the ligamentum flavum.

Both measurement techniques had unacceptable reliability (inter-rater ICC of 0.44 and 0.36 respectively) and discussion with observers suggested problems distinguishing the ligament from azygous joint capsule and lamina cortex, alongside difficulties selecting the correct site for the measurement. Due to this, no further investigation of flavum thickness was undertaken.

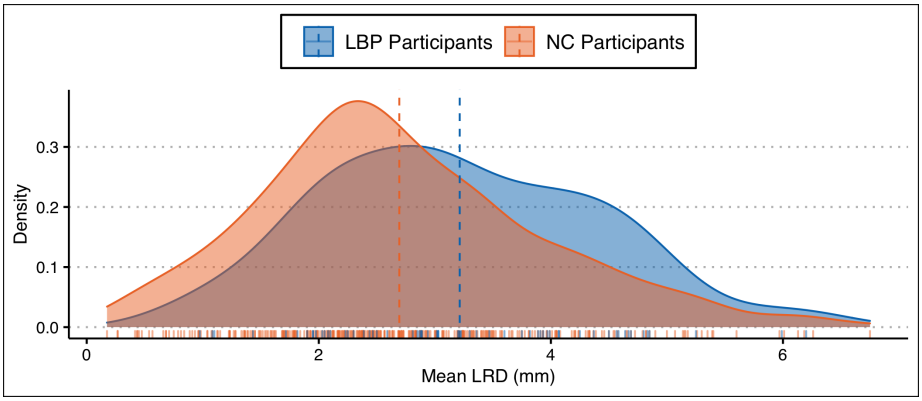


Figure 5.14: Distribution of mean lateral recess depth measurements per participant. Horizontal lines represent the mean value of the distributions.

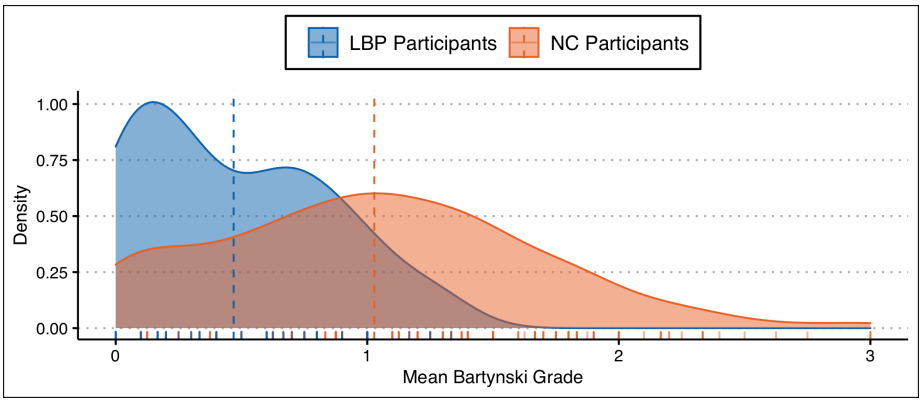


Figure 5.15: Distribution of mean Bartynski grades per participant. Horizontal lines represent the mean value of the distributions.

5.8 Chapter summary

Two main measurements of lateral recess were described: the lateral recess depth (LRD), a quantitative measurement of lateral recess narrowing, and the Bartynski grade, a grading system assessing qualitative features of transiting nerve root entrapment. Both measurements had good reliability. The lateral recess in those without qualitative features of canal narrowing was shown to progressively decrease in size between L1 and L4 before increasing again at L5 and the mean and minimum LRD were both negatively related to patient height and smaller in men.

As expected, the LRD and Bartynski grade were closely related, with a threshold LRD of 2 mm predicting the presence of qualitative nerve root entrapment with high sensitivity but poor specificity, the latter in part explained due to the ability of the transiting nerve root to escape compression by medial movement away from the narrowed lateral recess. Narrowing of the lateral recess and associated transiting nerve root entrapment were also closely related to narrowing of the central canal at the same level, unsurprising due to the common degenerative changes

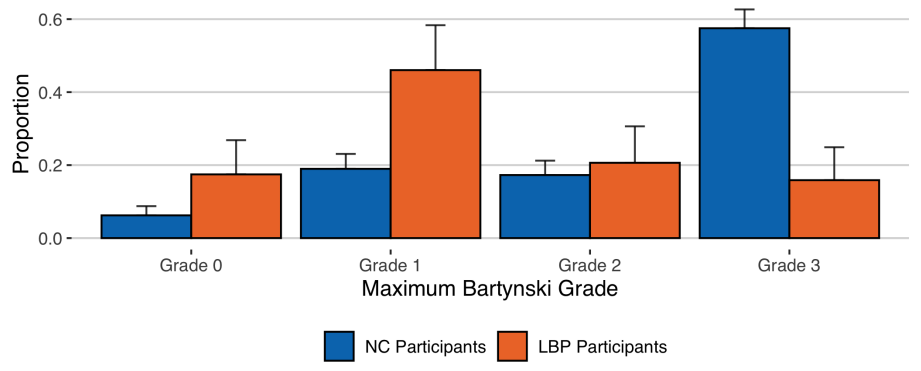


Figure 5.16: Distribution of maximum Bartynski grades per participant.

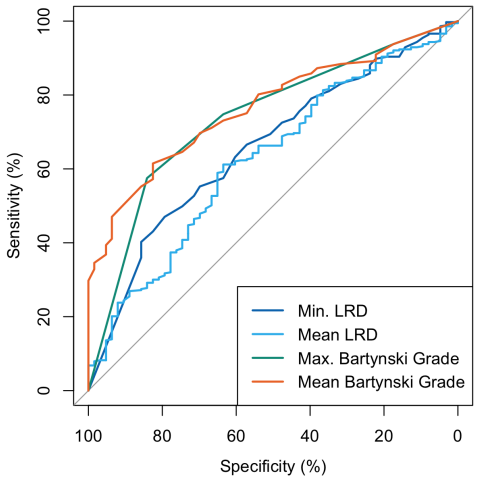


Figure 5.17: ROC curves demonstrating the ability of the investigated lateral recess measurements to separate NC and LBP participants.

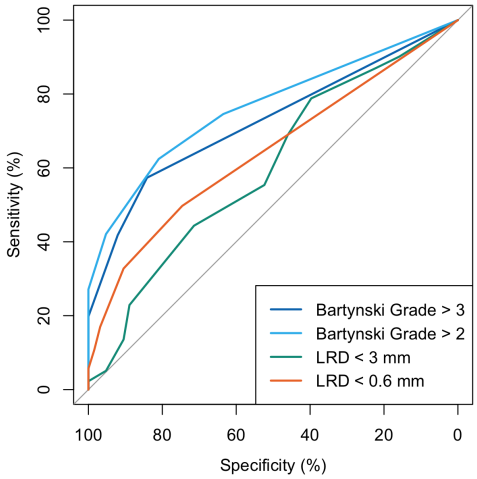


Figure 5.18: ROC curves for the ability of the number of stenotic lateral recesses, using various definitions of stenosis, to separate NC and LBP participants.

contributing to both lateral recess and central canal stenosis.

Narrowing of the lateral recess was most common at L3 and L4 and there was a weak inverse correlation between the minimum LRD per participant and patient age, a correlation which increased in strength when considering only patients with neurogenic claudication. The lateral recess was smaller, both at the minimum level per participant, but also on average across all levels per participant in NC compared to LBP participants. Similarly there were higher levels of qualitative nerve root entrapment in the lateral recess on NC claudication participants. Despite this, neither measurement separated NC and LBP participants well due to the significant overlap in measurement distributions between groups and there was no clear relationship between the measurements and clinical severity of symptoms in NC patients. Similarly, the number of levels judged as stenotic, using a variety of definitions, did not separate NC or LBP symptoms or predict severity of NC symptoms.

In addition to the measurements of lateral recess stenosis, the effect of significant disc herniations (defined as herniations causing deviation or compression of a nerve root) was assessed. Inter-rater reliability was much poorer for the presence of disc herniation compared to the other measurements discussed in this thesis. 32% of participants showed such herniations, which were most common at L5, and their incidence did not seem linked to demographic factors or the presence of neurogenic claudication symptoms.

Paradoxically, the presence of a significant disc herniation seemed to be linked to lower levels of neuro-ischaemic type symptoms in patients with NC. This finding was unexpected given the supposed pathological link between herniations and nerve root inflammation discussed in Section 1.6.2. There are a number of reasons to doubt the validity of this finding, the correlations identified are weak and p values are just under the threshold for significance. There was also no associated reduced disability with disc herniation, which would be expected with reduced pain levels.

Table 5.5: Various threshold measurements and their sensitivity and specificity for distinguishing between NC and LBP participant groups. Those measurements marked optimum are the point on the ROC curve with maximum Youden's J statistic.

Measurement	Threshold	Sensitivity (%)	Specificity (%)
Min. LRD (mm)	$<= 3$	90.3	15.9
	$<= 0.6$ (Optimum)	47.0	79.4
Mean LRD (mm)	$<= 2.8$ (Optimum)	61.2	63.5
Max. Bartynski	$= 3$ (Optimum)	57.5	84.1
	$>= 2$	74.8	63.5
Mean Bartynski	$>= 0.82$ (Optimum)	61.5	82.5

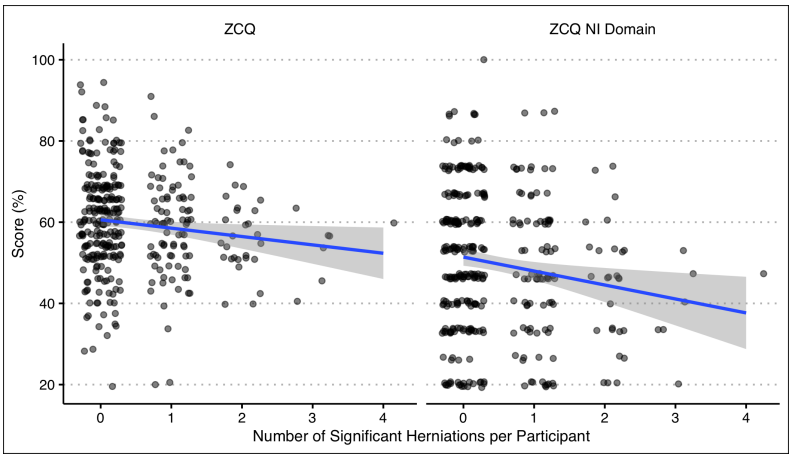


Figure 5.19: The relationship between the ZCQ symptom subscale and ZCQ neuroischaemic domain scores and the number of significant disc herniations per participant (those causing deviation or compression of a nerve root). Jitter has been added to reduce overlap between points.

Table 5.6: Comparison of clinical severity scores between patients with and without significant disc herniations (herniations displacing or compressing nerve roots).

Variables	Sig. Herniation Present				Sig. Herniation Absent				p Value
	Mean	CI			Mean	CI			
ODI (%)	31.0	28.7	–	33.3	33.5	31.9	–	35.2	0.252
ZCQ symptom-subscale (%)	57.8	55.9	–	59.7	60.6	59.1	–	62.1	0.034
ZCQ pain domain (%)	66.0	63.9	–	68.1	67.5	65.9	–	69.1	0.373
ZCQ neuro-ischaemic domain (%)	46.9	44.0	–	49.7	51.3	49.3	–	53.4	0.022
Walking distance to onset (m)	105.4	93.9	–	117.0	106.6	96.3	–	117.0	0.523
Walking distance total (m)	263.5	247.8	–	279.3	251.3	239.4	–	263.2	0.284
EQ-5D	0.70	0.67	–	0.73	0.67	0.64	–	0.69	0.221

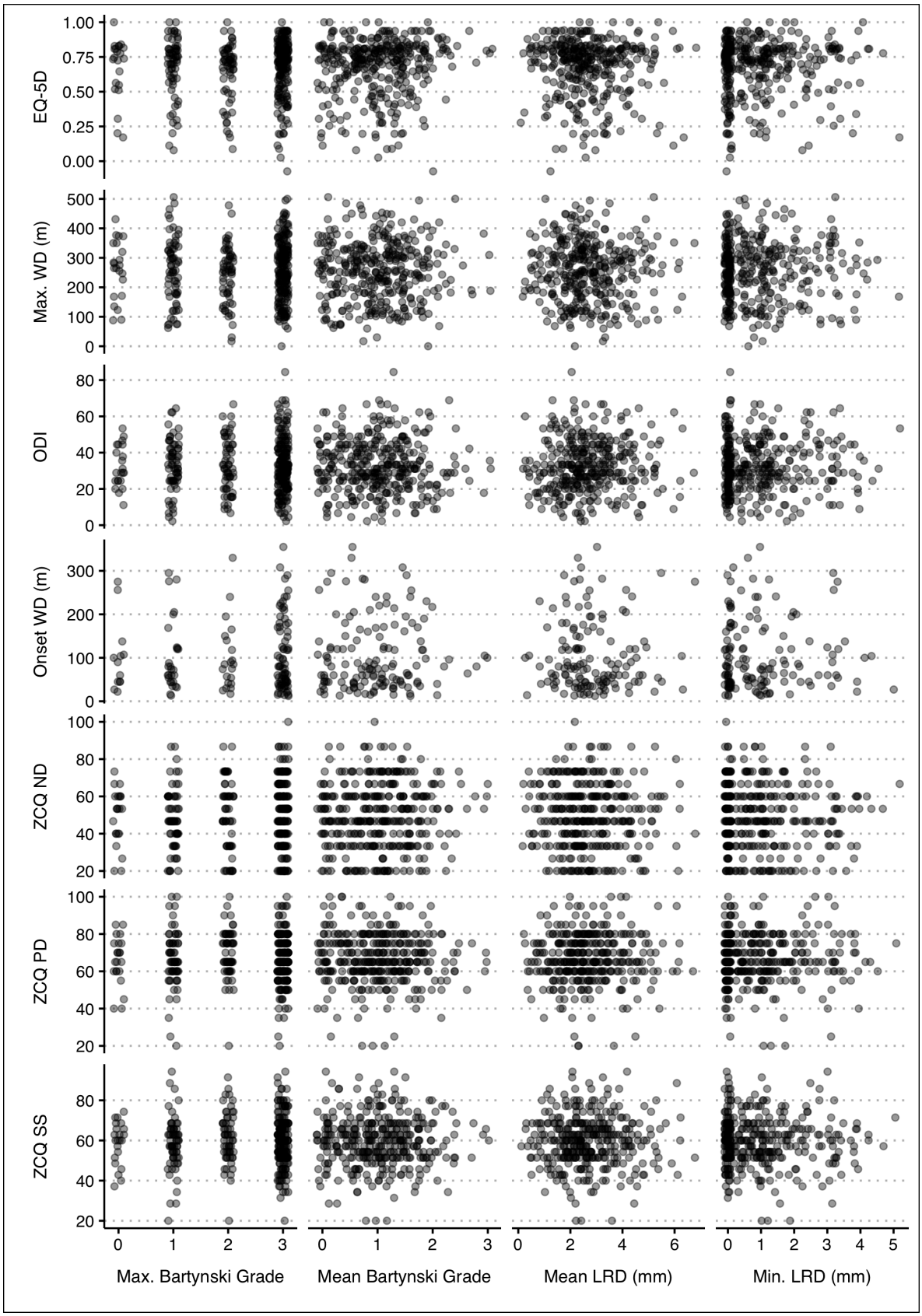


Figure 5.20: Scatter plots comparing measurements of the lateral recess to clinical severity scores for NC participants. Horizontal jitter has been added to reduce point overlap.

Chapter 6

The Neural Exit Foramen

6.1 Measurement method and reliability

The neural exit foramen diameter (NFD)

The neural exit foramen diameter (NFD) was a prespecified measurement for the BOOST sub-group analysis of treatment effect. Andreisek's core reporting criteria do not contain a quantitative measurement of neural exit foramen narrowing and hence one was selected based upon its common use in the literature (Steurer et al. 2011).

The measurement was made on axial T2 images cutting through the neural exit foramen (see figure 1.4, measurement 1). The axial slice was selected by comparing the axial image location data with sagittal images showing the same neural exit foramen: a sagittal slice was picked that clearly showed all the boundaries of the foramen, and an axial slice that cut through the exiting nerve root (as displayed on the sagittal image) was then selected. Once the image the measurement was to be performed on was selected, two measurement techniques were explored, one measuring the smallest diameter perpendicular to the canal and a second measuring at the most medial aspect of the exiting nerve root visible on the axial slice.

Inter-rater reliability was moderate, with ICC of 0.50 for the first technique and 0.40 for the second technique. The first technique was therefore selected and applied to the full imaging data-set where intra-rater reliability was found to be good with ICC of 0.67. Discussion between observers revealed difficulties with misregistration between sagittal and axial images when choosing the axial slice and consistent selection of the site of measurement within the visible canal on the axial image.

The Shapiro-Wilk test results was consistent with a non-normal distribution of the NFD ($W = 0.996$, $p < 0.001$), but the quantile-quantile (plot) (QQ) plot demonstrated the deviation from normal was minor; with evidence of a light lower tail, heavy upper tail and a minor left skew — figure 6.1. As such, parametric methods are used in the subsequent sections for analysis of the NFD.

The "Lee" Grade

The qualitative grading system proposed by S. Lee et al. (2010) aims to assess for entrapment of the exiting nerve root within a narrowed neural exit foramen. It is a 4 point grading system judged from primarily T1 weighted sagittal images at the site of most significant foraminal narrowing: grade 0 — normal foramina without evidence of nerve root compression; grade 1 – neural exit foramen narrowing results in perineural fat obliteration in two opposing directions without evidence of morphologic change in the nerve root; grade 2 — moderate neural exit foramen narrowing with obliteration of perineural fat in 4 opposing directions without evidence of morphologic change; grade 3 — severe neural exit foramen narrowing with evidence of nerve root collapse or morphologic change. Examples of different grades are given in figure 6.2.

The Lee grade was found to have a κ of 0.68 for both inter- and intra-rater reliability (87% and 92% agreement respectively), and far exceeded the reliability of other assessed measurements of foraminal nerve root entrapment, despite their inclusion in Andreisek's core reporting criteria (Andreisek et al. 2014) — see Section 6.7. Criticisms of the grading scheme from observers were very similar to the other trialled measurements, but less pronounced, and is discussed in later sections.

6.2 Changing neural exit foramen size with spinal level

There was a significant relationship between the NFD and spinal level, for both all measured neural exit foramina, and when only considering those without qualitative features of nerve root compression as judged by a Lee grade of 0 (ANOVA, $p < 0.001$ for both). The median NFD progressively increased between L1 and L3, decreasing slightly at L4, before again increasing at L5 — see figure 6.3.

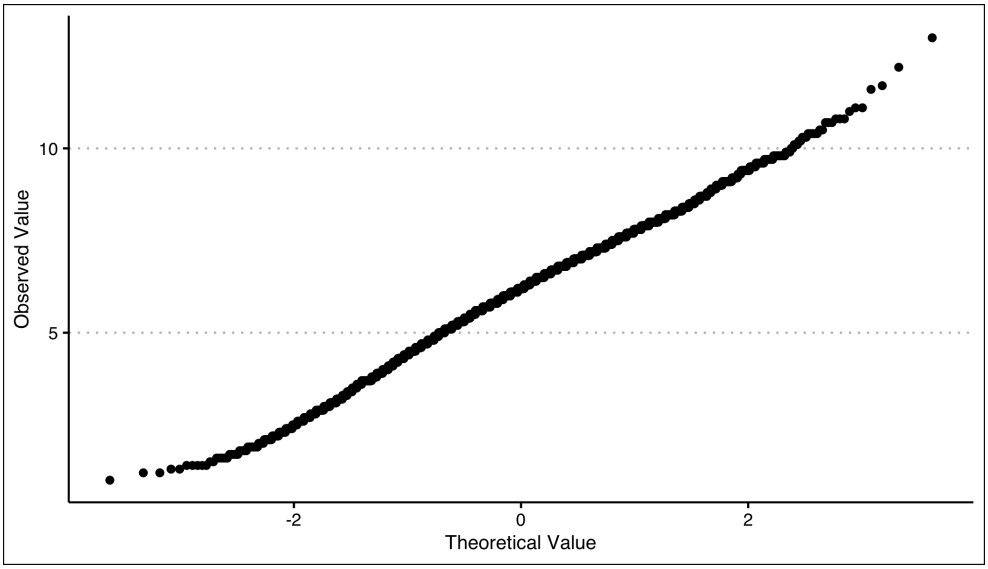


Figure 6.1: Quantile-quantile plot of NFD measurements.

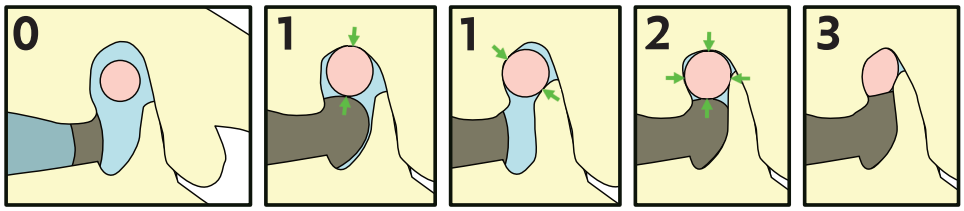


Figure 6.2: Schematic illustration of the Lee grade adapted from figure 1, S. Lee et al. (2010). Peri-neural fat is shown in blue surrounding the red nerve root. Two examples of grade 1 nerve root entrapment are given with differing directions of entrapment. In the grade 3 image the nerve root has lost its normal ovoid shape.

A neural exit foramen diameter of 3 mm is commonly used in the literature as a threshold below which the neural exit foramen is considered stenosed Andreisek et al. 2013. Only 4.6% of all measured foramina was considered stenotic by this definition, with the majority of such levels falling at L4 and L5 levels (see table 6.1). In contrast, 10% of graded levels had qualitative evidence of nerve root impingement (i.e. a Lee grade of 1 or above). Interestingly, despite the more common narrowing of foramen diameter at L4, the exiting nerve root was judged more commonly compressed at L5. One possible explanation for this effect may be the relatively oblique course the L5 neural exit foramen takes relative to the sagittal plane of imaging, potentially creating more artifactual compression via partial volume effects, but also potentially secondary to the longer neural exit foramen course (Wilmink 2010c).

6.3 Relationship between measurements

There was a moderate negative correlation between the NFD and the Lee grade ($R = -0.40$, $p < 0.001$), implying the NRD explains 16% of the variability in the qualitative grading. This correlation is smaller than that of the previously discussed relationships between quantitative and qualitative measurements of the central canal (where $R = 0.61$) and lateral recess (where $R = -0.64$). The distribution of NFD grades, broken down by associated Lee grades and spinal level is shown in figure 6.4a.

A ROC analysis was performed assessing the ability of the NFD to predict the presence of qualitative features of nerve root compression (i.e. a Lee grade of 1 or greater). The AUC was 77.6% (95% CI: 74.6% – 80.5%). Consistent with the lower correlation between the NFD and Lee grade compared to the equivalent correlations between qualitative and quantitative measurements in the lateral and central canal, AUC between NFD and Lee grade was significantly smaller than the corresponding AUC between LRD and Bartynski grade (Dunn's test, $p < 0.001$).

The commonly used 3 mm LRD threshold for neural exit foramen stenosis had a 99.4% specificity, but only 33.7% sensitivity for observer recorded nerve root entrapment. In comparison, the optimal threshold for detecting entrapment, as determined by maximising Youden's statistic, was 4.7 mm, which reduced the specificity to 86.8% but increased the sensitivity to 61.1%.

There was a clear relationship between LRD meas-

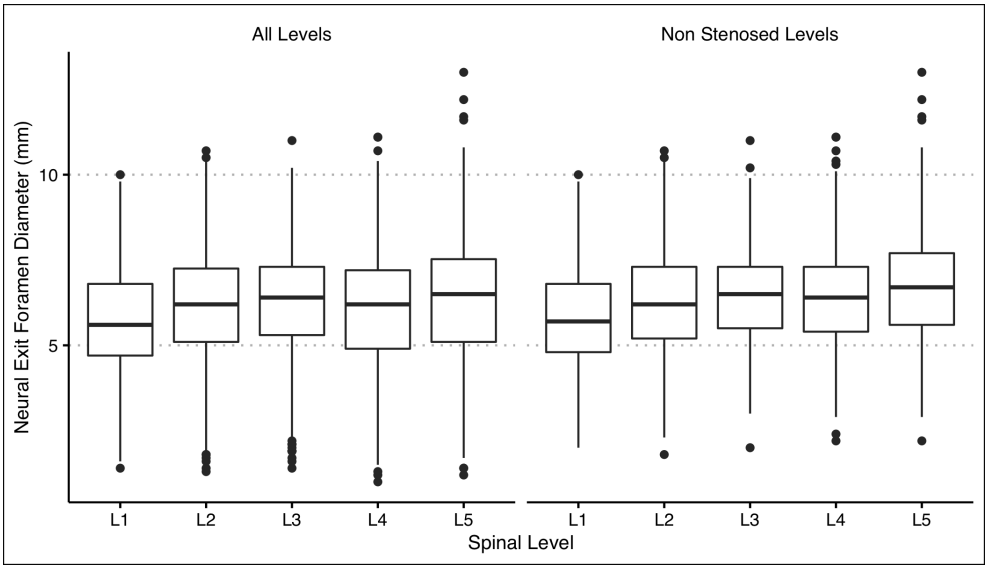


Figure 6.3: The distribution of neural exit foramen diameter with changing spinal level. Graphs for both all measured levels and only those without qualitative features of foraminal nerve root compression are shown.

Spinal Level	NFD $<$ 3 mm (%)	Lee Grade $>$ 0 (%)
L1	2.9	1.6
L2	3.6	3.4
L3	3.6	8.4
L4	6.1	15.5
L5	5.6	21.2

Table 6.1: Sites of neural exit foramen stenosis as defined by a NFD < 3 or the presence of qualitative features of nerve root compression.

urements and measurements of the central canal at the same spinal level for each individual (see figure 6.5a and 6.5b). No significant correlation between APD and NFD was identified at L5, though significant correlations were identified at all other levels. Similarly, the correlation between DS-CSA and the LRD was weaker at L5 than at the other lumbar levels.

As expected, given that facet joint hypertrophy/osteophytosis contributes to narrowing both the lateral recess and neural exit foramen, there was a moderate correlations between the NFD and LRD (figure 6.5c).

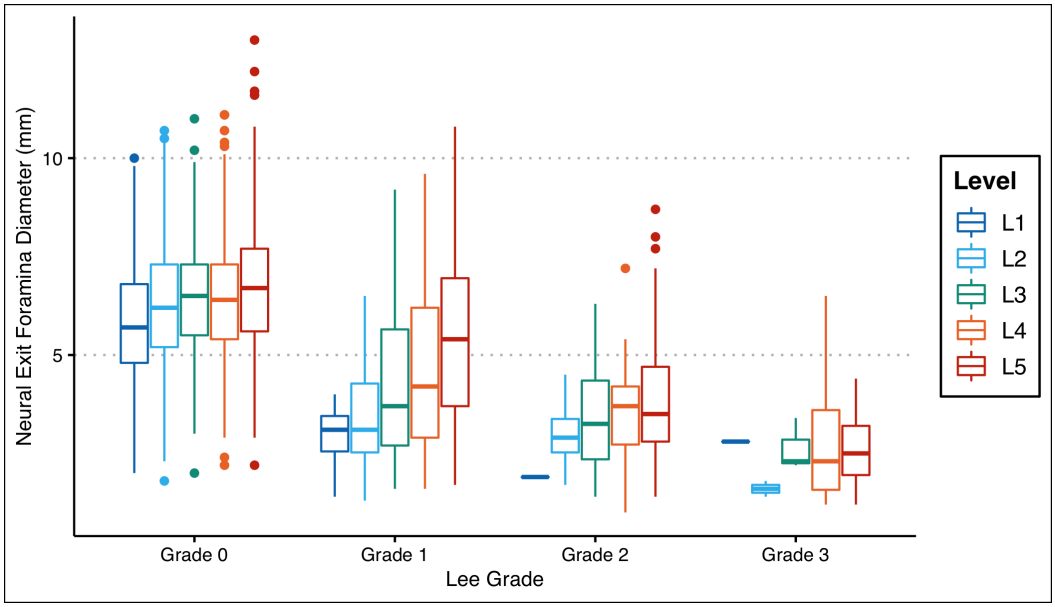
The correlations between the NFD and quantitative measurements of the central canal and lateral recess were weaker than the corresponding correlations between the LRD and measurements of central canal discussed in sections 4.3 and 5.3. They also demonstrated much less pronounced effect of spinal level on the gradient of the line of best fit (see Section 5.3).

There were also significant relationships between the risk of nerve root impingement within the neural exit foramen (i.e. a Lee grade > 0) and the risk of impingement in the lateral recess at the same level and side of the canal (i.e. a Bartynski grade > 1) or the central canal at the same level (i.e. a Schizas grade > B). The odds ratio for the presence of impingement was 2.22 (95% CI: 1.73 – 2.85) and 2.51 (95% CI: 2.03 – 3.11) when comparing the central canal and lateral recess to the neural exit foramina respectively ($p < 0.001$ for both).

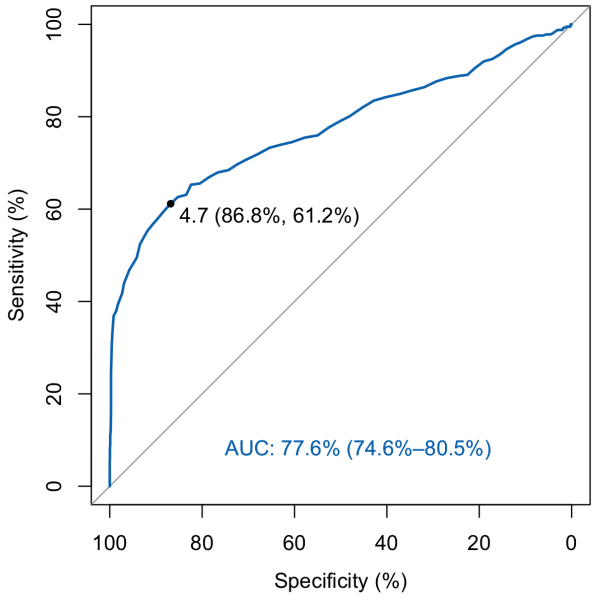
6.4 Relationship to demographic variables

A weak negative correlation between the minimum NFD per participant and age was identified (figure 6.6) without an associated significant correlation with the mean foraminal diameter per participant. There was a similar weak positive correlation between the maximum Lee grade per participant and participant age (figure 6.7), again without an associated correlation with the mean Lee grade per participant.

No significant correlation was identified between the mean or minimum NFD or mean or maximum Lee grade per participant and participant BMI or height. There was also no difference in the neural exit foramen measurements comparing between gender groups, ethnicity groups or between those patients with a preexisting clinical MRI study compared to those having a MRI for the purposes of the BOOST trial.

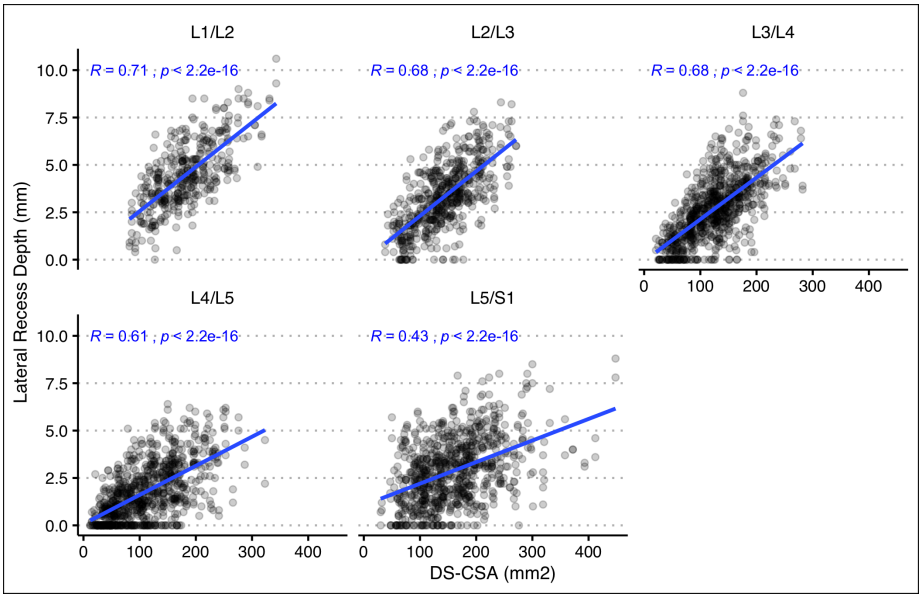


(a) Box-plots demonstrating the relationship between the NFD and Lee grade, broken down by spinal level.

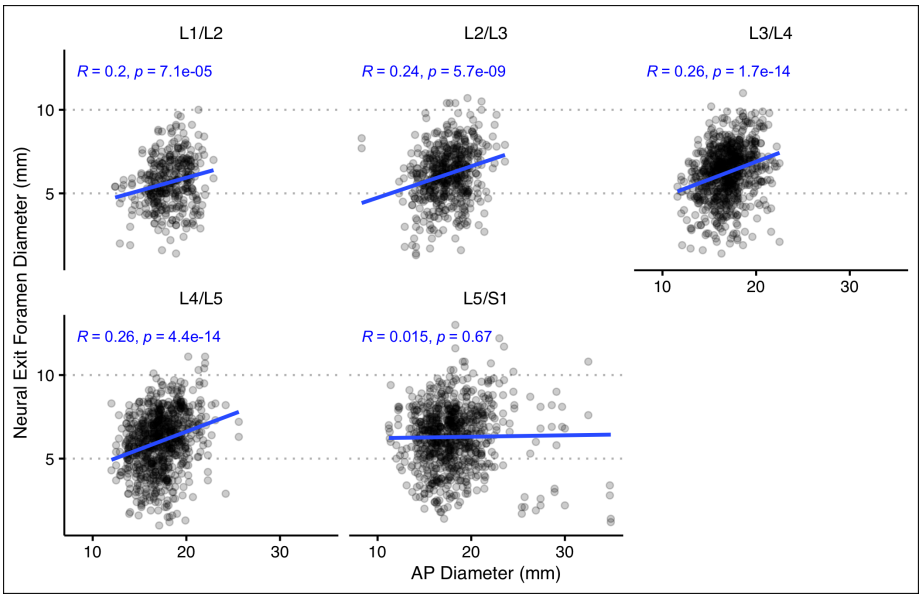


(b) ROC analysis: the ability of the NFD to predict a Lee grade of 1 or above.

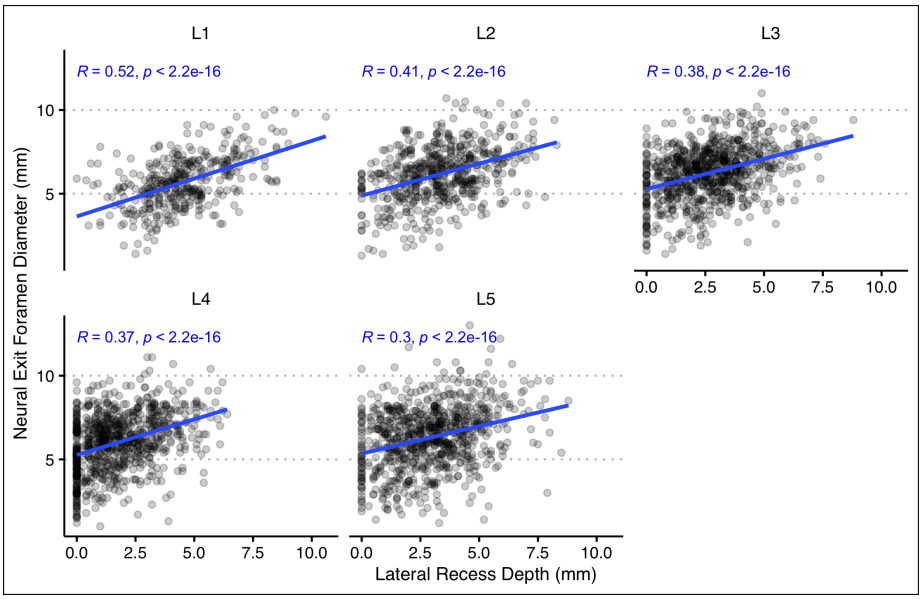
Figure 6.4: Assessing the relationship between NFD and Lee grades.



(a) Scatter plots of NFD and DS-CSA.



(b) Scatter plots of NFD and APD.



(c) Scatter plots of NFD and LRD.

Figure 6.5: The relationship between NFD and other measurements of canal size.

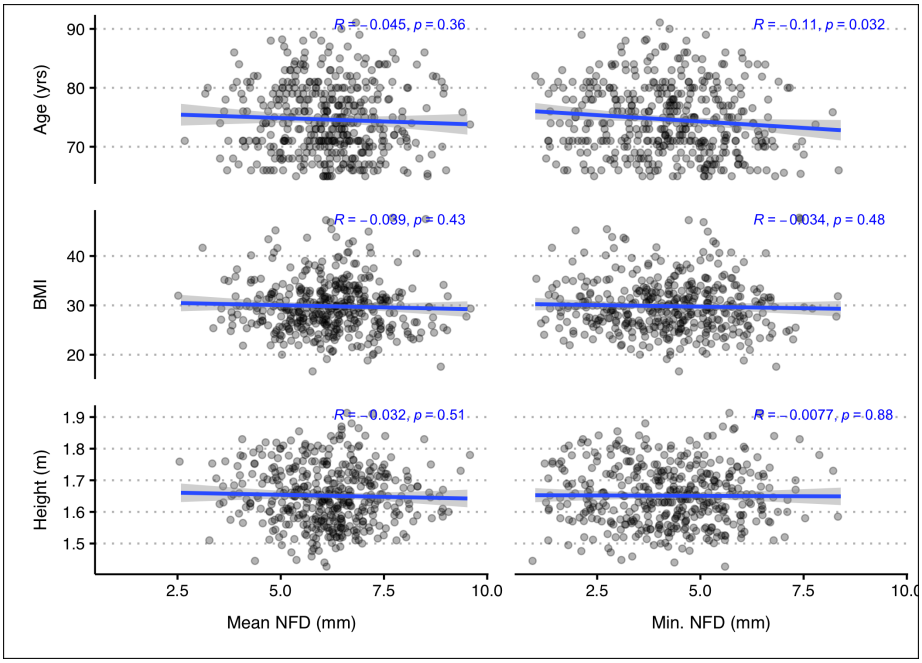


Figure 6.6: The relationship between mean and minimum NFD per participant and participant age, BMI and height measurements.

6.5 Differences between NC and LBP imaging groups

Both the mean and minimum NFD was smaller in NC participants compared to LBP participants (table 6.2) but, in a similar fashion to that seen for the lateral recess, there was substantial overlap in their distributions (figures 6.8a and 6.8b) with poor AUC values (figure 6.9a) indicating they did not accurately separate NC and LBP participants. There was no significant difference in AUC values between the mean and minimum NFD.

Similarly, higher mean and maximum Lee grades were seen in NC participants compared to LBP participants ($p < 0.001$ and $p = 0.008$ respectively). Lee grades of 2 and 3 were relatively rare across all included participants, but were seen more frequently in neurogenic claudication patients (figure 6.8c). Again, due to the degree of overlap in distributions, both measurements had poor AUC values. The AUC for the mean Lee grade per participant was 0.60 (95% CI: 0.54 – 0.66) and 0.60 for the maximum Lee grade per participant (95% CI: 0.54 – 0.66). The AUC for the mean NFD per participant was significantly higher than the AUC values for the mean and minimum Lee grades (Dunn's test, $p < 0.016$ for both), but there was no significant difference when comparing to the AUC for the minimum NFD per participant.

The performance of a variety of diagnostic thresholds for neural exit foramen stenosis are demonstrated in table 6.3. The 3 mm NFD threshold commonly used in the literature had a relatively high specificity for neurogenic participants but poor sensitivity. The optimum threshold of the maximum Lee grade per participant, as found by Youden's J statistic, was 2 or greater, which also had a high specificity but poor sensitivity. The overall accuracy for all thresholds across all measurements was poor.

To investigate the role of multi-level neural exit foramen stenosis, the ability of the number of neural exit foramina recesses meeting various thresholds of narrowing per spine were assessed for their ability to distinguish LBP participants from NC participants (figure 6.9b). There was no significant improvement in AUC values using this technique compared to the performance of the measurements discussed above. In fact, the AUC for mean NFD per participant was significantly greater than the AUC for the number of stenotic levels using all the tested thresholds for defining stenosis (Dunn's test, $p < 0.03$ for all).

6.6 Relationship to symptom severity in NC participants

No significant correlation was identified between any clinical severity variable and the mean or minimum LRD per participant or the mean or maximum Lee grade per participant (figure 6.10). Similarly there was no correlation between any clinical variable severity and the number of stenotic neural exit foramina counted using the same thresholds employed in figure 6.9b and no significant difference between single and multi-level neural exit foramina stenosis groups using the same thresholds.

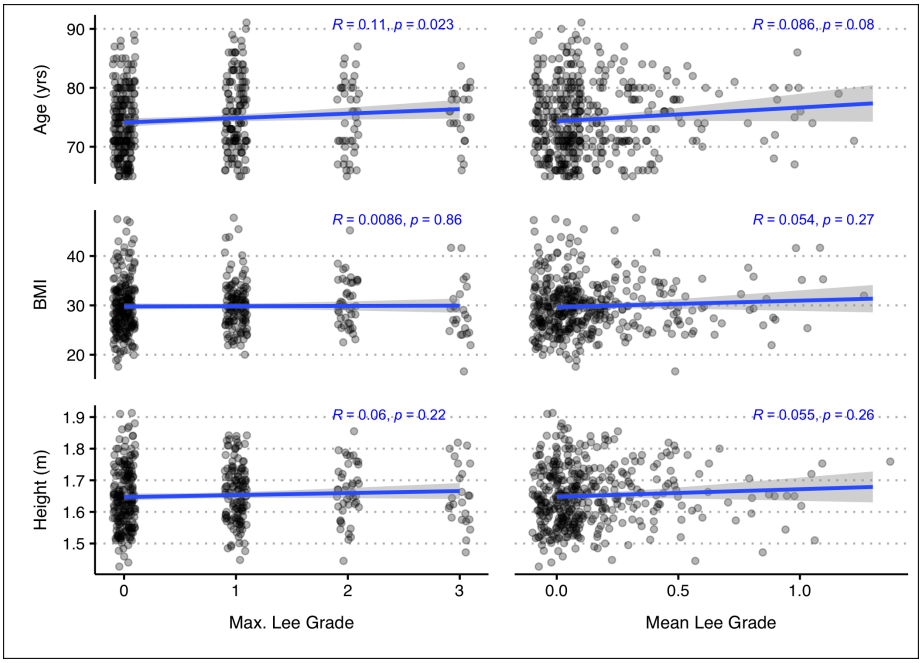


Figure 6.7: The relationship between maximum and mean Lee grades per participant and participant age, BMI and height.

Table 6.2: Comparison of neural exit foramen measurements between NC and LBP participant groups.

Variables	NC Participants				LBP Participants				p Value	ROC Analysis			
	Mean	CI			Mean	CI				AUC	CI		
Min. NFD (mm)	4.11	3.96	–	4.25	5.05	4.71	–	5.39	< 0.001	0.67	0.60	–	0.74
Mean NFD (mm)	6.01	5.89	–	6.12	6.92	6.64	–	7.2	< 0.001	0.72	0.65	–	0.79
Mean Lee Grade	0.15	0.12	–	0.17	0.06	0.03	–	0.09	< 0.001	0.60	0.54	–	0.66

6.7 Other measurements

The "Lurie" grade

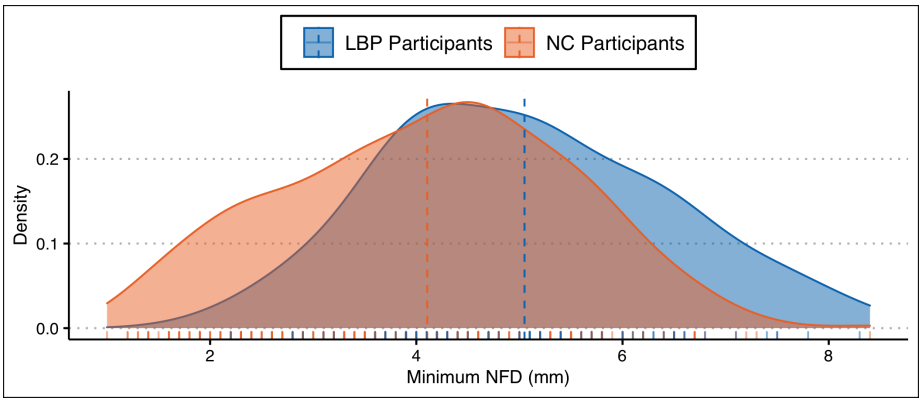
This grading system, proposed by Lurie et al. (2008a), is the same as the Lurie grade for central canal stenosis (discussed in Section 4.7) but applied to the neural exit foramen. It is a formalisation of the normal, mild, moderate or severe stenosis descriptions common to radiology reports and was suggested by Andreisek as a core reporting criteria for compromise of the foraminal zone (Andreisek et al. 2014). Similar to the problems found when applying the Lurie grade to central canal stenosis, the grading system was found to suffer from severe problems with reliability with inter-rater κ of 0.17 (54% agreement). As such, no further investigation using this grading system was performed.

The "Wildermuth" grade

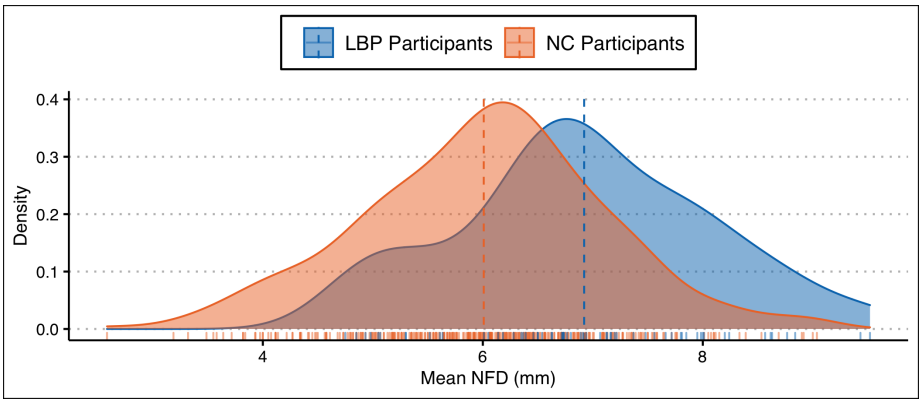
The grading system proposed by Wildermuth et al. (1998) is a 4 point qualitative grading system aiming to assess exiting nerve root compression within the neural exit foramina. The four grades are defined as: 1 — normal foramina size and form of the epidural fat; grade 2 — slight stenosis with deformity of the epidural fat but which still completely surrounds the exiting nerve root; grade 3 — marked foraminal stenosis with epidural fat partially surrounding the nerve

Table 6.3: Various threshold measurements and their sensitivity, specificity and accuracy for distinguishing between NC and LBP participant groups. Those measurements marked optimum are the point on the ROC curve with maximum Youden's J statistic.

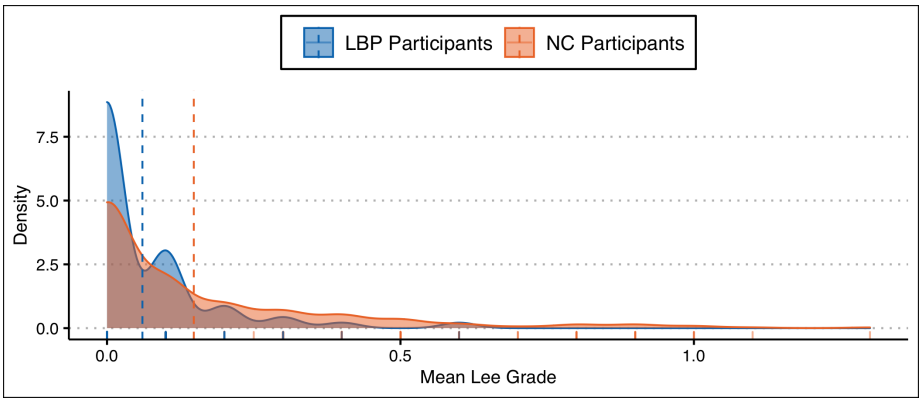
Measurement	Threshold	Sensitivity (%)	Specificity (%)
Min. NFD (mm)	<= 3.65 (Optimum)	36.8	88.9
	<= 3	24.4	93.7
Mean NFD (mm)	<= 6.50 (Optimum)	68.8	68.3
Maximum Lee Grade	>= 2 (Optimum)	18.6	98.4
Mean Lee Grade	>= 0.15 (Optimum)	30.8	87.3



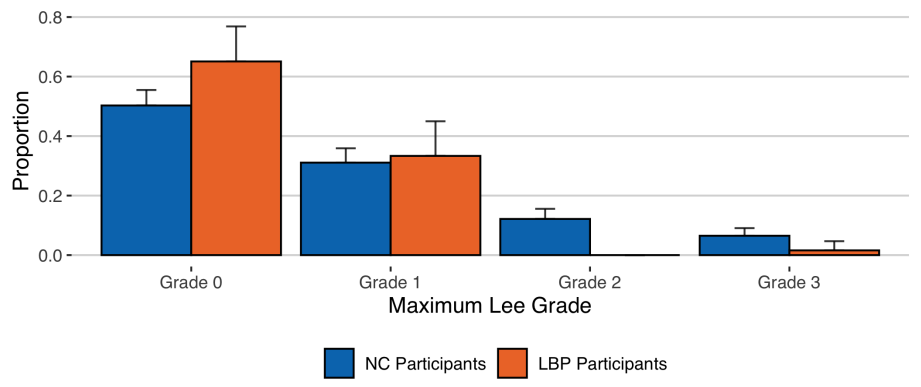
(a) Distribution of minimum NFD measurements per participant. Horizontal lines represent the mean value of the distributions.



(b) Distribution of mean NFD measurements per participant. Horizontal lines represent the mean value of the distributions.

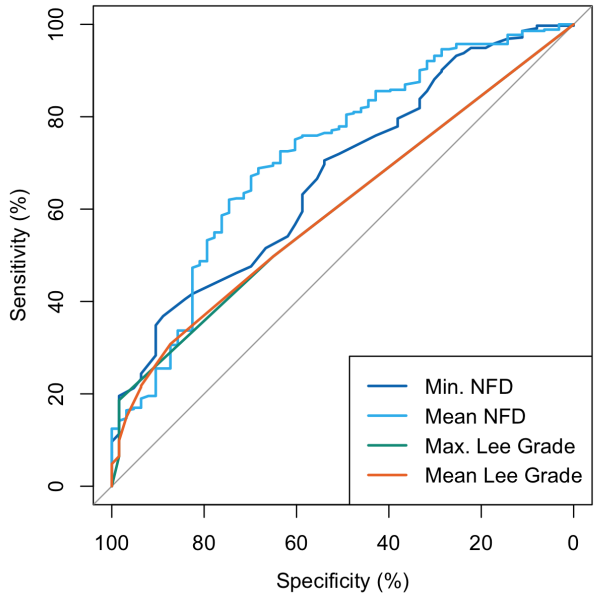


(c) Distribution of mean Lee grades per participant. Horizontal lines represent the mean value of the distributions.

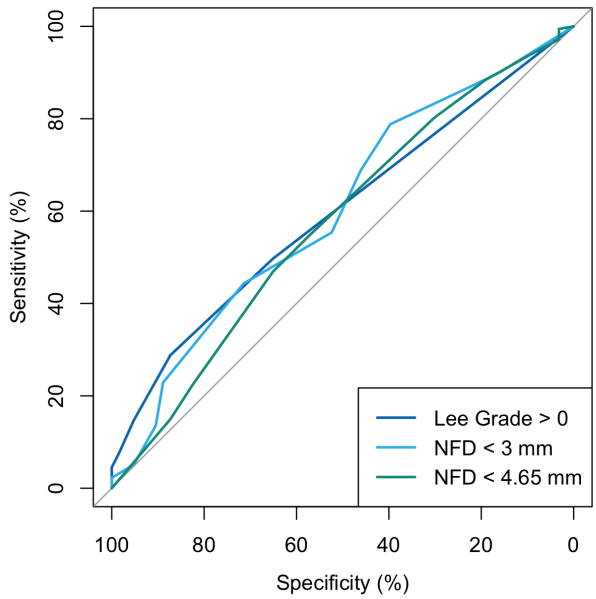


(d) Distribution of maximum Lee grades per participant.

Figure 6.8: Distribution of neural exit foramen measurements in NC and LBP patient groups.



(a) ROC curves demonstrating the ability of the investigated neural exit foramen measurements to separate NC and LBP participants.



(b) ROC curves for the ability of the number of stenotic neural exit foramina, using various definitions of stenosis, to separate NC and LBP participants.

Figure 6.9: ROC Analyses

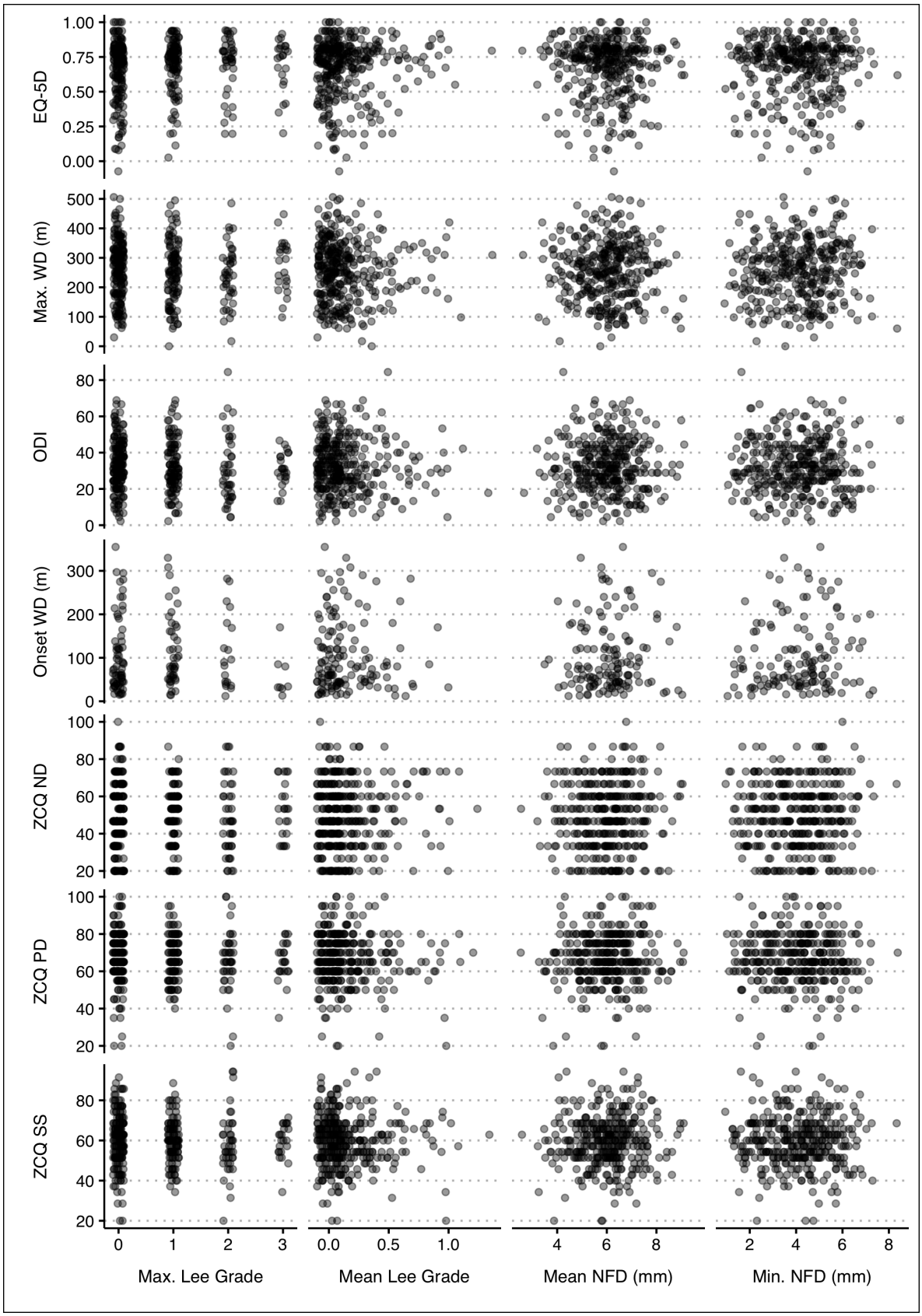


Figure 6.10: Scatter plots comparing measurements of the lateral recess to clinical severity scores for NC participants. Horizontal jitter has been added to reduce point overlap.

root; grade 4 — advanced stenosis with complete obliteration of the epidural fat. It was chosen in the first round of measurement assessments over the grading system proposed by Pfirrmann et al. (2004), and endorsed by Andreisek's as a core reporting criteria, as the former was specifically designed for use in the degenerate neural exit foramen, whereas Pfirrmann's grading system relates more to disc herniation related nerve root entrapment outside the central canal.

The overall reliability of the Wildermuth grading system was moderate, with a inter-rater κ of 0.48 (87% agreement). Discussion between observers and manual inspection of discrepancies noted problems with consistent assignment of grade 2 stenosis, due to difficulties defining epidural fat deformity consistently, and problems with potential grade 3 foramina, due to nerve roots often brushing against the boundaries of otherwise completely normal foramina where no impingement would be expected. It was often also difficult to decide whether a thin rim of epidural fat remained between a potentially impinging structure and the nerve root. Due to these issues, it was decided to instead use the Lee grading system described in Section 6.1 — which does not include a Grade 2 or grade 3 equivalent, and instead only records impingement of the nerve root between 2 or more opposing surfaces.

6.8 Chapter summary

Two main main measurements of the neural exit foramina were described: the neural exit foramen diameter (NFD), a quantitative measurement of neural exit foramen narrowing, and the Lee grade, a 4 point qualitative grading system assessing for exiting nerve root entrapment. The former had only a moderate inter-rater reliability while the Lee grading system proved more reliable in terms of its inter- and intra-rater reliability. Both measurements were correlated, but this relationship was weaker than the corresponding relationships between quantitative and qualitative measurements of the central canal and neural exit foramina, with the variation in LRD only explaining an estimated 16% of the variability of the Lee grade.

Neural exit foramen stenosis, defined by a threshold level of foramen narrowing, was most common at the L4 level, but when instead defined by the presence of qualitative features of exiting nerve root entrapment, was more common at L5. The latter finding being potentially secondary to the longer more oblique foramen course relative to the more superior levels. The degree of neural exit foramen narrowing was related to both central canal and lateral recess narrowing at the same spinal level, as expected given the related degenerative changes contributing to narrowing of all three structures. Similarly, the presence of qualitative features of exiting nerve root entrapment was predictive for features suggesting nerve root entrapment in both the lateral recess and central canal at the same spinal level.

The neural exit foramen was smaller on average in patients with neurogenic claudication (NC) compared to those with lower back pain (LBP), both when considering only the smallest foramen per spine or when averaging across all neural exit foramina per spine. There was a corresponding higher rate of exiting nerve root entrapment in NC patients. Despite this the measurements of the neural exit foramina failed to accurately separate NC from LBP patients and there was no identified relationship between the degree of neural exit foramen narrowing, exiting nerve root entrapment or the number of stenosed foramina and clinical severity of NC symptoms.

Chapter 7

Lumbar Spinal Alignment and Fractures

7.1 Spondylolisthesis and spondylolysis

Measurement

For all disk levels between L1 and L5, the degree of spondylolisthesis (anterior or posterior translation of a vertebral body relative to the level below) was graded on T2 weighted sagittal images using the system proposed by Meyerding (1932). The system grades the degree of translation on a 5 point scale as either: grade 0 – no translation relative to the size of the end-plates at that level, grade I – 0-25% translation, grade 2 – 25-50% translation, grade 3 – 50-75% translation or grade 4 – 75-100% translation. Translation less than 3 mm was graded as grade 0 (Kauppila et al. 1998). Both inter and intra-rater reliability was good with κ of 0.76 (96.3% rater agreement) and κ of 0.80 (94.9% rater agreement) respectively.

In total, 48.7% of all included participants had spondylolisthesis affecting at least one spinal level. Spondylolisthesis was most common at the L4/5 and L5/S1 levels, where 29.7% and 20.1% of levels demonstrated at least grade 1 spondylolisthesis respectively. Participants with spondylolisthesis were slightly older (mean 75.5 years vs 73.8 years, Mann-Whitney $p = 0.001$), and slightly taller (mean height 1.66 m vs 1.64 m, Mann-Whitney $p = 0.004$). There was no significant difference in BMI between those with and without spondylolisthesis and no significant difference in maximum spondylolisthesis grade between gender or ethnicity groups.

Within participants of the BOOST RCT, participants who had a pre-existing clinical MRI, as opposed to a MRI for the purposes of the trial, were much more likely to have spondylolisthesis – 58.4% of participants with a pre-existing MRI showed spondylolisthesis, as opposed to 43.2% without (Chi-squared $p = 0.008$).

In total, 7.4% of participants had at least one level with a raised APCR, a measurement strongly associated with spondylolysis. Spondylolysis represents the presences of bilateral breaks in the pars interarticularis which allows translation of the vertebral body and posterior neural arch relative to each other. Spondylolysis was more common in the lower lumbar spine, with 6.8% of L5 and 1.2% of L4 levels affected. Spondylolysis was associated with spondylolisthesis at 50% of levels. These numbers likely underestimate the true number of levels with spondylolysis, as spondylolysis without translation of either the neural arch or vertebral body will not have a raised APCR.

Patients with spondylolysis were on average taller than those without (1.70 m vs 1.65 m, Mann-Whitney $p = 0.001$) and more likely to be male (14.2% of men had spondylolysis as opposed to 2.1% of women, Chi-squared $p = 0.004$). There was no significant difference in age, BMI or ethnicity group between those with or without spondylolysis and no difference in BOOST participants between those with and without a pre-existing clinical MRI study.

Relationship to canal size

The relationship between the degree of spondylolisthesis and the central canal size was related both to spinal level and the presence of co-existing spondylolysis at that spinal level.

At L1 and L2, where very few levels showed spondylolisthesis there was no identified relationship between spondylolisthesis grade and DS-CSA. In contrast, at L3 and L4 where many more levels showed spondylolisthesis, its presence was associated with smaller DS-CSAs (Man-Whitney, $p < 0.001$, table 7.1).

Spinal levels with evidence of spondylolysis showed significantly higher DS-CSA measurements compared to those without — 201 mm^2 (CI: 177 – 226 mm^2) vs. 138 mm^2 (CI: 135 – 140 mm^2, Mann-Whitney < 0.001) and this complicated the relationship between spondylolisthesis and canal size at L5 where the incidence of spondylolysis was highest. The overall relationship between DS-CSA and spondylolisthesis at L5 was non-significant, but became significant when levels with spondylolysis were excluded, again showing smaller DS-CSA measurements at spondylotic levels (Man-Whitney, $p = 0.003$).

No significant relationship between the APD and spondylolisthesis was identified between L1 and L4. APD measurements at L5 levels showed larger APDs in levels with spondylolisthesis compared to those

Table 7.1: Comparison of DS-CSA between levels with and without spondylolisthesis. P values from Man-Whitney U tests are given.

Level	Spondylolisthesis Absent				Spondylolisthesis Present				p values
	Mean DS-CSA (mm^2)	CI			Mean DS-CSA (mm^2)	CI			
L1	182.18	176.88	–	187.48	175.06	155.17	–	194.94	0.55
L2	144.52	139.64	–	149.40	135.5	114.93	–	156.07	0.48
L3	128.50	123.42	–	133.58	95.23	79.88	–	110.58	< 0.01
L4	128.27	121.64	–	134.90	74.51	66.24	–	82.78	< 0.01
L5	161.86	154.82	–	168.89	148.36	134.25	–	162.47	0.08
L5 (No Spondylolysis)	160.55	153.49	–	167.61	132.94	120.37	–	145.51	0.00

without (Man-Whitney, $p < 0.001$) but once the effect of spondylolysis had been excluded the relationship became insignificant.

The presence of spondylolisthesis was also associated with significantly smaller lateral recess diameters at all lumbar levels (Mann-Whitney, $p < 0.045$) and smaller neural exit foramina diameters at L2, L3 and L5 (Mann-Whitney, $p < 0.038$). The neural exit foramina diameters at L1 and L4 did not significantly differ between those levels with and without spondylolisthesis. As expected, the presence of spondylolysis was associated with larger neural exit foramina diameters (Mann-Whitney $p = 0.05$) but was not associated with a significant difference in neural exit foramina diameter.

Relationship to clinical variables

The distribution of maximum spondylolisthesis grades per participant differed significantly between NC and LBP groups (table 7.2, Chi-squared $p = 0.048$) with an AUC of 0.56 (95% CI: 49.6 – 63.2%). Overall 50.8% of NC participants had at least one lumbar level with spondylolisthesis, as opposed to 36.5% of LBP participants. The ability of the maximum spondylolisthesis grade per participant to differentiate NC from LBP participants was very poor, with AUC of 0.56 (95% CI: 0.50 – 63.0).

In contrast, there was no difference in the incidence of spondylolysis in NC or LBP groups and there was no significant difference in any clinical severity score between participants grouped by the presence of spondylolisthesis or spondylolysis.

7.2 Scoliosis

Observers rated each spine for the presence of significant lumbar scoliosis. Significant scoliosis was defined as as maximal Cobb angle of greater than 20° between any two lumbar vertebra. As no formal coronal imaging was performed the judgement was made based upon coronal scout sequences.

Reliability was perfect with all observers agreeing across all observations. 10% of all participants showed lumbar scoliosis. Scoliosis was more common in women (12.8% of female participants as opposed to 6.6% of male participants, Chi-squared $p = 0.046$). There was no significant relationship to patient age, BMI, height or ethnicity, and no difference in prevalence between BOOST participants with a pre-existing MRI compared to those having an MRI for the purposes of the trial.

The median values of the minimum DS-CSA, mean DS-CSA, mean APD were significantly larger in individuals with scoliosis compared to those without (see table 7.3). This may in part be explained geometrically, with cross-sectional of the dural-sac elongating as the canal becomes more angulated relative to the plane of imaging. In contrast, the minimum NFD was smaller in patients with scoliosis, though measurement was often difficult, again due to angulation. There was no significant difference in minimum LRD between those with and without scoliosis and no relationship with the presence of spondylolisthesis.

The presence of scoliosis was not related to NC or LBP status and did not show a significant relationship to any clinical severity variable among NC participants.

7.3 Fractures

MRI observers recorded all levels showing evidence of osteoporotic vertebral body compression fractures on sagittal images. Fractures were identified by loss of vertebral body height compared to adjacent levels.

Reliability was good with inter-rater κ of 0.72 (97.6% agreement) and inter-rater κ of 0.73 (97.8% agreement).

Unsurprisingly for this age group, 17% of individuals had at least one vertebral compression fracture. Fractures were most common at T12 and L1 levels where 9.3% and 8.2% of levels were fractured respectively. There was no relationship between the presence of fractures and patient age, BMI, height, gender, ethnicity, or, for BOOST participants, whether the participant had a pre-existing scan or one done spe-

Table 7.2: Comparison of maximum spondylolisthesis grades in NC and LBP participants

Spondylolisthesis Grade	None	Grade 1	Grade 2
Neurogenic claudication participants (%)	49.2	48.0	2.8
Low back pain participants (%)	63.5	31.8	4.8

Table 7.3: Comparison of spinal canal size between those with and without scoliosis. P values from Man-Whitney U tests are given.

Variables	Scoliosis Present				Scoliosis Absent				p Value
	Mean	CI			Mean	CI			
Min. DS–CSA (mm^2)	105.1	90.6	–	119.6	87.6	82.9	–	92.2	0.022
Mean DS–CSA (mm^2)	154.3	139.7	–	168.8	137.0	132.7	–	141.2	0.031
Min. APD (mm)	16.1	15.5	–	16.7	15.7	15.6	–	15.9	0.163
Mean APD (mm)	18.2	17.7	–	18.7	17.5	17.3	–	17.6	0.010
Min. LRD (mm)	1.5	1.0	–	1.9	1.0	0.9	–	1.1	0.060
Min. NFD (mm)	3.6	3.2	–	4.0	4.3	4.2	–	4.5	0.001

cifically for the purposes of the RCT.

The presence of fractures was related to increased pain and disability, with significantly higher ZCQ pain subdomain scores, reduction in walking distance to symptom onset, reduction in the maximum distance the patient could walk, and a reduction in the EQ-5D (see table 7.4). The ODI was also greater (indicating greater disability) in individuals with fractures but the difference did not reach significance. Interestingly there was no significant difference in the overall ZCQ symptom subscale due to a lack of difference in items contributing to the ZCQ neuro-ischaemic subdomain, suggesting the associated increase in pain was not related to the patients claudication symptoms. There was no significant difference in the presence of fractures between NC and LBP participants.

7.4 Chapter summary

This chapter discussed spinal alignment and osteoporotic compression fractures within the included participants.

Spondylolisthesis was very common, being present in just under half of all participants and with increased prevalence in older and taller participants. Spondylolisthesis was generally associated with smaller central canals and lateral recesses, but this relationship was complicated in those individuals with isthmic spondylolisthesis, the latter instead associated with larger spinal canals. Spondylolisthesis, but not spondylolysis, was more common in NC patients, but had little ability to differentiated them from the control group, and neither were related to clinical severity within individuals with neurogenic claudication.

Scoliosis was present in 10% of participants and more common in women. It was associated with larger canal sizes compared to non-scoliotic participants, most likely due to elongation of the canal cross-section as the canal becomes angled relative to the plane of imaging. Scoliosis did not appear to be related to NC symptoms or symptom severity within NC participants.

Fractures were also common within the included participant groups, affecting 17% of participants. The presence of a fracture was associated with higher disability, lower maximum walking distance, shorter walking distances to symptom onset and overall increased pain scores (but without significant differ-

Table 7.4: Comparison of clinical severity variables between individuals with and without osteoporotic compression fractures T12-L1 in patients with neurogenic claudication.

Severity Variables	Fracture Absent				Fracture Present				p Value
	Value	CI			Value	CI			
ODI	32.2	30.7	–	33.7	35.4	32.1	–	38.6	0.067
ZCQ symptom subscale	59.1	57.9	–	60.4	62.4	59.3	–	65.5	0.107
ZCQ pain subdomain	66.3	65.0	–	67.7	70.6	67.4	–	73.8	0.026
ZCQ neuro–ischaemic subdomain	49.5	47.7	–	51.3	51.4	47.1	–	55.8	0.423
WD to symptom onset	112.9	103.9	–	121.9	74	61.1	–	86.8	0.032
Max. WD	261.4	251.2	–	271.7	223.7	199.7	–	247.7	0.005
EQ–5D	0.69	0.67	–	0.71	0.60	0.55	–	0.66	0.025

ence in the claudication type pain, and fractures were not predictive of claudication symptoms). Fractures were hence was one of the few MRI features identified within this study clearly related to symptom severity, highlighting the likely multi-factorial nature of a patients presentation.

Chapter 8

Combining Anatomical Measurements

8.1 Introduction

Both the previous literature (see Chapter 2) and work presented within this thesis up to this point have largely focused on the minimum and mean of a single given measurement of canal size for the diagnosis of anatomical LSS. This approach has the benefit of simplicity but suffers from a number of limitations.

The minimum of a given measurement of canal size only gives information about one point in the canal, ignoring other sites of degeneration and potential nerve root compression. The mean, in contrast, is a course statistic which gives information about the general size of the canal, but which may be misleading about the potential for nerve root compression given the propensity of canal degeneration to be markedly more pronounced at certain levels and sites while relatively minor at others (see § 1.4.3).

The use of single measurements to determine canal size also ignores the potential sites of compression at which that measurement does not apply. For instance, using only the DS-CSA to decide if a spine is stenotic fails to consider potential potential nerve root compression in the lateral recesses or neural exit foramina. This failure is particularly egregious given the proposed pathophysiological mechanisms involving multi-site nerve root compression in the generation of neurogenic claudication symptoms (§ 1.6.3).

As previously described in § 1.4.2, individuals also vary in terms of the size of their normal spinal canal (prior to onset of degenerative changes). If narrowing of the spinal canal relative to the individuals normal (pre-degeneration) canal size is of greater pathological significance than the absolute degree of narrowing, raw measurements, without normalisation to the patients body size, will fail to accurately reflect the anatomical changes of significance.

Finally, in this current work we have recorded several different ways of assessing for potential nerve compression at each anatomical site assessed. Combining the information from these different measurements has the potential to improve the overall accuracy of assessment of nerve root compression and hence the accuracy of the prediction of symptomatology as a whole.

The following section presents a number of approaches designed to address some of the above limitations and hopefully improve upon the heretofore poor diagnostic performance of the explored canal measurements.

8.2 The stenosis ratio

8.2.1 Rationale

Laurencin et al. (1999) originally described the stenosis ratio (SR) in 1999. This was the only measurement of lumbar canal narrowing found during the scoping review presented in § 2.1 that made any attempt to relate canal narrowing to the patients expected canal size unaffected by degenerative change.

The stenosis ratio at a given spinal level consists of the disc level DS-CSA divided by the mid-pedicular DS-CSA, the idea being that the mid-pedicular level is not significantly affected by degenerative change and hence should act as a marker of the normal size of the canal (see § 1.4.3). Hence patients without significant degenerative change at the disc level are expected to have a SR measurement of 1 with progressive stenosis giving lower values. Schizas et al. (2010) found that the Laurencin's SR was smaller in patients with clinical LSS when compared to a control group, and was predictive of failure of conservative LSS treatment (though see Chapter 2 for criticism of the methodology of such papers).

Unfortunately direct measurement of the stenosis ratio was not possible for the majority of participants within the current study — the MRIs largely have axial images taken in blocks centred around the disc level. This frequently leaves gaps in imaging coverage at the level of the pedicles and hence prevents measurement of the pedicular DS-CSA. This is likely

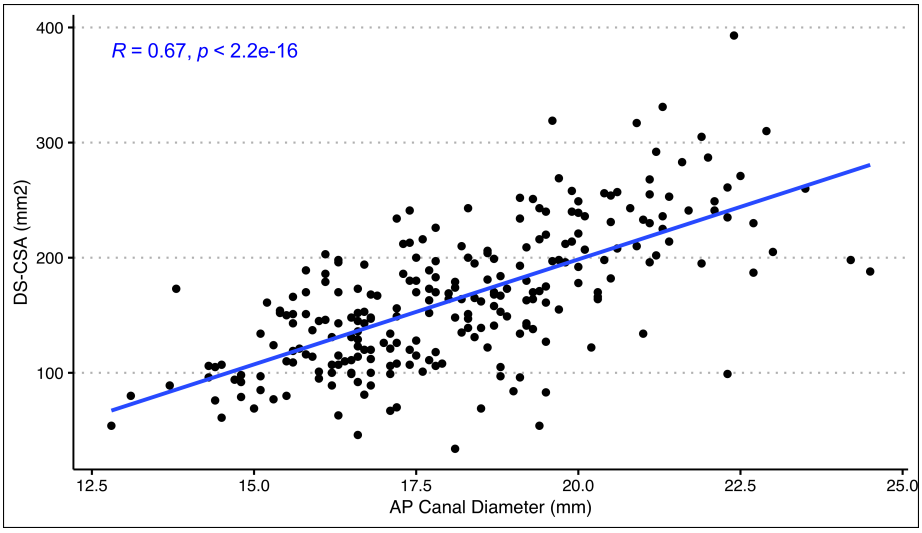


Figure 8.1: The linear fit between the APD and DS-CSA in LBP patients

to be the case for most MRI studies performed in the NHS, which largely follow a similar imaging protocol.

The antero-to-posterior canal diameter (APD) is available within the study data set and has the potential to act as a normalising measurement for the DS-CSA. Like the pedicular DS-CSA, it is measured away from the motion segment and hence away from the majority of degenerative changes. It is largely normally distributed (deviating from normal in those with spondylosis) and does not vary significantly with the presence of qualitative features of stenosis or correlate with the patients age (§ 4.3). It is weakly correlated with patient height ($R = 0.1$) allowing it to act as a weak marker of body size (§ 4.4). While the small number of levels with spondylosis potentially have sizes unrelated to the non-degenerate size of the canal due to the incomplete neural arch, their effect on the overall utility of the measurement should be small due to the relative rarity of involved levels, the fact that spondylosis was generally associated with increased spinal-canal cross-sectional areas rather than stenosis, and the preponderance of spondylosis at L5, a level more rarely associated with spinal stenosis (§ 7.1).

While a simple ratio of the DS-CSA to the APD might seem therefore to be a potential simple solution to the lack of pedicular DS-CSA, it has a number of issues as a which make it unsuitable, the largest of which is mathematically undesirable properties due to the combination of different measurements. For instance, changes in APD would have little effect on the ratio compared to changes in the DS-CSA due to the relative smaller range of the APD measurement within the study population. Such a ratio would also not have an easily intuitive interpretation.

Instead, a "non-degenerate" DS-CSA measurement for the participant and spinal level can be potentially predicted from the APD. The actual DS-CSA can then be divided by the predicted DS-CSA to give a measurement that behaves and is interpretable in a manner similar to Laurencin's stenosis ratio.

The process of predicting the expected DS-CSA without degenerative change is made easier by the fairly good linear relationship between the DS-CSA and APD measurements within in the data set, as previously presented in figure-4.5, allowing the predictions to be made from a linear model fitted to this data — producing a predicted SR. Ideally a "normal" population without symptoms change would be used to fit the linear model for the predictions, as the population within the study are unlikely to be representative of the general population, due to their symptomatology, with an over-representation of stenotic levels. This is partially mitigated by the inclusion of all measured levels — degeneration tends to be a process that affects one motion segment while sparing others, and exclusion of neurogenic claudication patients, who are known to have smaller canals, from the data used to build the predictive model. The lack of a definitely normal population on which to base the predictions remains a significant limitation of this technique and should be born in mind when assessing the results below.

Another way to think about this transformation of the DS-CSA is in terms of a dimensionality reduction. We know that the DS-CSA correlates fairly well with the Schizas grade but also with the APD which does not correlate with the Schizas grade (§ 4.3). By removing the variation in the DS-CSA that can be explained by the APD the residual variation may more closely relate to the degree of degeneration, rather than the variation in non-degenerate canal size, which is represented by the APD. Because of this, we would expect the strength of correlation between the maximum Schizas grade and minimum predicted SR to be stronger than that seen between the minimum DS-CSA.

8.2.2 Results

A linear model was fitted between the DS-CSA and APD, via the method of least squares, for all available spinal levels in the low-back pain control group. Two levels with isthmic spondylolisthesis and outlying

Table 8.1: Comparison of the predicted stenosis ratio in NC and LBP participants.

Variable	NC Participants		LBP Participants		p Value
	Value	95% CI	Value	95% CI	
Min. Ratio	0.57	0.54 – 0.60	0.74	0.69 – 0.80	< 0.001
Mean Ratio	0.92	0.89 – 0.94	0.99	0.94 – 1.03	0.005

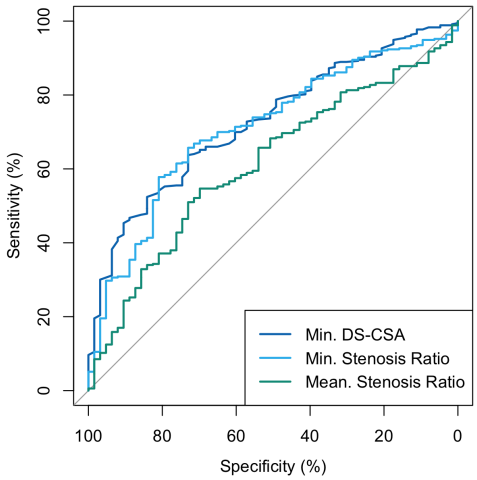


Figure 8.2: ROC analysis for the predicted stenosis ratio.

APD measurements were excluded from the model. The model had a β of 18.3, an intercept of -166.4 and a adjusted R^2 of 0.45 (figure 8.1). Hence the DS-CSA could be predicted using the following formula:

$$\widehat{DSCSA} = APD \times 18.3 - 166.4 \qquad (8.1)$$

The predicted SR was calculated for each measured level, followed by the mean and minimum ratio per canal. The minimum predicted SR significantly negatively correlated with the maximum Schizas grade ($R = -0.73$, $p < 0.001$) and the absolute strength of correlation was stronger than that seen between the maximum Schizas grade and the unaltered minimum DS-CSA ($R = 0.52$, § 4.3).

Both the mean and minimum SRs were significantly smaller in NC participants compared to LBP participants (table 8.1). Despite this the predicted SR did not effectively differentiate LBP and NC participants with no significant improvement in AUC compared to the to the un-adjusted minimum DS-CSA per participant (Figure 8.2). Neither mean nor minimum predicted SR had any significant correlation with clinical severity.

8.3 General machine learning classifiers

An earlier version of the work within this section has been published as a conference abstract — Gagen, RM et al. (2020)

8.3.1 A brief introduction to machine learning

Machine learning techniques allow automated created of "classifiers" — algorithms trained to predict the classification of an object based upon other data associated with that particular object and its classification in general (Witten 2017). One of the classic teaching examples, used within the machine learning literature, is training an algorithm to classify 3 different species of Iris flowers based upon their petal and sepal length and width (Fisher 1936). In an example of supervised learning, the machine learning algorithm of interest is fed a data-set of flowers of known species alongside their associated measurements (the training-set) and the generated algorithm is then tested for accuracy of species prediction, on new data the algorithm has not been exposed to during training (the test-set).

Such algorithms have great potential applications for medical diagnostic work due to their ability to combine information from multiple sources into making a diagnosis based upon data from previous cases, in a similar way to which a human diagnostician might work. This approach stands in contrast to the single variable based receiver operator characteristic analysis so far performed in this thesis and allows data from all collected MRI measurements as well as demographic data to be combined in the attempt to produce a model separating NC from LBP patients, without reliance on summary variables such as the minimum or mean of a given measurement.

A good example of a machine learning classifier

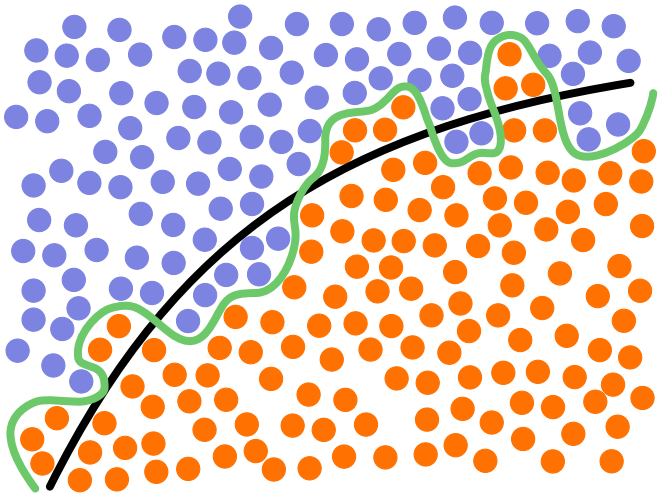


Figure 8.3: The figure illustrates two models designed to separate orange and blue data-points represented by the black and green lines. The green line represents an over-fitted model. It has perfect accuracy on this particular set of data but is likely to have less overall accuracy than the regularised black line when applied to new data points from the same population.

is a decision-tree (p. 105, Witten 2017). Numerous decision-tree based classifiers exist, and their full workings are beyond the scope of this introduction, but in general they work by repeatedly dividing the data-set to produce a branching flow chart like structure, with a question based on some attribute of the data at each node of the tree. To predict the classification of an object you start at the root node, and answer each question in turn based on the data available for that object, following the branches indicated by the answer until at the end of the branch is a prediction of that objects classification.

The decision-tree is built during model training by selecting the variable within the training-set that best classifies the objects in the training data based on some optimum threshold value of that variable. The training-data is then divided based upon that threshold and the same process is then performed on each new sub-division (branch) of the training data. This process is continued until either the all of the new sub-divisions of the data contain only objects of one classification, or until some threshold of tree complexity is reached. Many decision-tree algorithms then go on to perform a pruning step which removes excessive branches that do not contribute to the overall tree accuracy (p. 214, Witten 2017).

Aside from decision tree based algorithms numerous different forms of classifier exist, ranging from simple rule based classifiers to complex neural networks, however selection of which algorithm to use for a given problem is not straightforward. Increasing the complexity of the algorithm selected does not always equal an improvement in performance on a given task and often simple algorithms may perform as well or better than complex algorithms (Holte 1993). Given this, selection of an algorithm is often a case of trying several and selecting the one that works best on the task at hand. This process is further complicated by the need to select hyper-parameters for each algorithm — these are configuration variables external to the model which are not selected from the data and include such things as variables specifying how aggressively a decision tree will be pruned after training (p. 171, Witten 2017).

In general, all work towards model selection and tuning of hyper-parameters should be performed using the training data. Only once the final model and hyper-parameters are selected are they evaluated on a separate test-data set that is only used for final model accuracy evaluation.

The need for this strict separation of training and test stages comes from the risk of over-fitting the model and, in association, over-estimating the models likely real world performance. Over-fitting is a process by which a given model comes through training to over-represent the particular features of the training data rather the general trends also present in the population from which the training data is drawn (Everitt et al. 2010). This process is illustrated in figure 8.3. An over-fitted model will have high accuracy when applied to the training data, but when exposed to new data will show a fall in prediction accuracy, sometimes dramatically.

The risk of over-fitting increases with the number of variables in the model, use of automated variable selection procedures, and pre-testing of multiple models built in different ways (Babyak 2004). Training a model when there are multiple unknowns — such as which variables to select for the model or what specific machine learning technique would be best to use — often becomes a balancing act between exploration of the effects of different parameters on model accuracy and avoidance of model over-fitting.

Interim assessment of model accuracy to allow model selection and hyper-parameter tuning before use of the test set can be achieved using a technique known as k-fold validation (where k is an integer value). k-fold validation involves shuffling the training-set, dividing the shuffled training-set into k equally sized groups and then looping through each group. Each group is then used as a test-set for a model trained on the remaining $k - 1$ groups (p. 181

James et al. 2013). This procedure produces k sub-models and their accuracy can then be summarised as an estimate of the accuracy of a model trained with the same parameters on the full training data. This final model trained on all available data is expected to perform better than the sub-models due to the greater size of data available for training. k is most commonly set to 10 (p. 184, James et al. 2013).

It is essential to avoid repeated testing of multiple different models on the final test-set as this again leads to overfitting, this time to the test data itself, rather than to the training data, and is likely to lead to an over-estimate of model accuracy which cannot be detected without collection of new data to form a new test set.

8.3.2 General approach

This section sets out the general approach taken to training the machine learning models described in the result sections.

Variable selection

The large number of variables collected from the MRI studies for each participant made a single large machine learning model impractical due to the risk of overfitting (Babyak 2004) and long computing time required to train the algorithms. In addition, inclusion of variables with little predictive power may reduce model accuracy by interfering with the learning process by increasing the dimensionality of the data without a corresponding increase in its efficacy for prediction (Verleysen et al. 2005). Hence, an initial decision was made to restrict the number of variables used in model training.

In previous chapters, the majority of variables not directly related to canal size (such as the presence of fractures, herniation or measurements of spinal alignment) did not significantly differ between NC and LBP participants. This suggests they would be unlikely to contribute to prediction accuracy and were therefore excluded from the model. Demographic data, while potentially be of use in placing the canal measurements in context, were also removed due to the minor differences in demographic features between between NC and LBP groups (discussed in 3). These differences most likely occurred secondary to sampling error and are hence not representative of real world differences in these populations. Inclusion of these variables could therefore spuriously increase the estimate of model accuracy relative to their likely real world performance. In addition, all previously discussed data on presence of multi-level stenosis or the number of stenotic levels per spine were also excluded on the basis of its general failure to improve on classification accuracy over the measurements at the most severe level.

A decision was also made to split the approach into two separate strategies: one training models based upon the summary measurements, for example the mean and minimum DS-CSA per canal, and one based upon the raw measurements alone without inclusion of summary measurements. This approach kept the number of variables in any one model smaller, and avoided duplication of measurement values within the data for each participant. The first strategy will allow assessment of the value of combining information about the most severe sites of stenosis. The second strategy has, with a complex enough model, the potential to effectively derive the variables present in the first strategy and information on the presence of multi-level stenosis without this data being manually calculated and provided.

Before training, the variable set for each strategy was screened for near-zero variance variables — those with one or very few unique values relative to the number of samples. Such variables can cause machine learning (ML) algorithms to crash or to produce unstable model fits (p. 4, Kuhn 2008). No problematic variables identified were for among the variables for the first strategy, but the variables for the second strategy contained five variables with near-zero variance: the Schizas grade at L5/S1, and the Lee grade at both L1 and L2 bilaterally. These variables were removed from the data-set.

The MRI data set was split into a training and test-set based upon a 75:25 ratio, with stratification of the split based on the LBP/NC grouping to ensure the ratio of LBP to NC participants remained similar in both training and test-sets. Ten cross-validation folds were created from the training data and this process was repeated 3 times to produce a total of 30 different folds for interim assessment of model accuracy. Repeating fold creation with averaging of model performance across a larger number of folds gives a more robust model assessment than using a single folding procedure (p. 70, Kuhn et al. 2013).

Algorithm selection

Four machine learning algorithms were selected for evaluation chosen to cover a variety of common machine learning approaches.

A classification and regression tree (CART) based algorithm was selected (Therneau et al. 2019). Decision trees have many benefits, including the ability to handle both categorical and continuous data, with little requirement for pre-processing to correct skewness or other non-normal characteristics of the data. Decision trees can also be inspected in a human readable way, allowing direct understanding of how the decisions are being made, unlike many other ML algorithms which effectively function as a black box (a systems whose workings are not readily understood) (James et al. 2013). Decision trees similarly allow analysis of the relative importance of the different variables in the input data in terms of final classification (Greenwell et al. 2020).

The second model chosen was a random forest based model: ranger (Wright et al. 2017). Random forest classifiers are ensemble classifiers that work by constructing a multitude of decision trees during the training stage, each based upon slightly different ver-

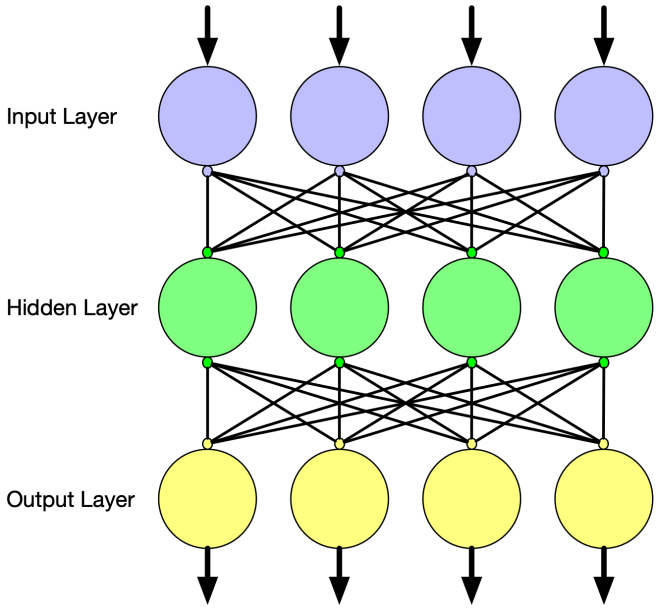


Figure 8.4: Representation of the layers of a multi-layer perceptron.

sions of the training data. The output classifications from the algorithm are produced by combining the decisions of individual trees. This approach to model construction often overcomes the tendency to overfit to the data seen with algorithms based upon individual decision trees (p. 587, Hastie et al. 2001) while retaining many of the benefits listed in the paragraph above. Random forest based algorithms are typically more robust to changes in hyper-parameters than individual decision trees further reducing the risk of overfitting.

A penalised logistic regression model was also selected — the glmnet algorithm implemented by Friedman et al. (2010). Binary logistic regression is quick to train, often effective and is powered by the multiple linear regression equation expressed in logarithmic terms. (p. 314, Field et al. 2012).

Finally, a multi-layered perceptron was chosen, a form of feed-forward neural network, specifically TensorFlow (Allaire et al. 2020b), an algorithm originally developed by Google®, and accessed using the Keras application programming interface (API) for R (Allaire et al. 2020a).

The full workings of neural networks is outside the scope of this thesis (and frankly outside my understanding) but roughly the network consists of multiple layers of interconnected nodes which function in a manner somewhat like neurons, with nodes being activated by input from the layer behind and giving output to the layer in front (Witten 2017, p. 420,). An input layer of nodes allows data to enter the network, and an final exit layer ends the network and emits classifications. The hidden layers between the input and output layers are trained by back-propagation of feedback from each classification, and this feedback over the course of training epochs shapes each nodes response to its inputs (figure 8.4). The great advantage of such networks is that they can potentially distinguish data that is not linearly separable (Cybenko 1989), but their disadvantage is their black-box like nature – it is difficult if not impossible to understand why a trained neural network is making its classifications. Due to the significant computation time taken to train such an algorithm on the hardware available to me, neural network were trained as part of the second strategy only, where the more complex data set was thought to be a better fit for a neural networks potential strengths.

For each algorithm a range of possible hyper-parameters were chosen (described further within the results sections) with models trained for each possible combination of hyper-parameters using the cross-validation folds. The AUC achieved by each model was averaged across all training folds for each model/hyper-parameter combination. The standard error for each mean was also calculated based on variations in performance across the training folds. The best combination of hyper-parameters was then chosen for each algorithm by selecting the least complex model across all hyper-parameter settings which achieved an AUC within one standard error of the best AUC achieved in the training stage. This technique of hyper-parameter selection was described by Breiman (1998) and is designed to reduce model complexity (and potential over-fitting) with minimal compromise of model performance.

Data pre-processing

All necessary pre-processing of data was performed using parameters selected based only on the training data, both within cross-folds and also on final model training before application to the test-set. Strict avoidance of pre-processing steps based on the whole data set is essential to prevent over-estimation of model accuracy by preventing data leakage. Data leakage occurs when information about the test-set, which would not be available in a real world applica-

Measurement	Min./Max. Measurement AUC	Mean Measurement AUC
APD	0.62	0.63
DS-CSA	0.73	0.69
Schizas Grade	0.70	0.71
LRD	0.66	0.63
Bartynski Grade	0.74	0.77
NFD	0.67	0.72
Lee Grade	0.60	0.60

Table 8.2: A summary of AUCs for the ROC curves of the various measurements of spinal canal size, summarised per participant as either the measurement at the spinal level most severely narrowed or the mean across all available measurements for that participant.

tion, becomes included in the training-set (Kaufman et al. 2011). For example when normalising data for model input, if the standard deviation and mean for the normalisation were taken from the whole data set, rather than just the training set, information on the standard deviation and mean of the test set would leak into the training set when the normalisation was applied.

The pre-processing steps for all algorithms involved correcting for the imbalance in the number of LBP to NC participants. Such an imbalance compromises the processes of model training, as the model will often tend to focus on the most prevalent classification, optimising accuracy by simply predicting everything to be the majority class (p. 64, Witten 2017). To avoid this the random over sampling examples (ROSE) method described by Menardi et al. (2014) was adopted. This is a way to reduce class imbalance by random oversampling from the minority class. This step was only performed during training and was not applied to the final test-set to any avoid potential distortion of final accuracy assessment.

One of the assumptions of logistic regression is that all input variables are non-colinear and independent (p. 342, Field et al. 2012). The variables within the data-set for both strategies show marked inter-variable correlation and are not truly independent. Generally this problem can be overcome by a strategy of either "feature elimination" or "feature extraction". The former consists of manually selecting features for removal from the data-set, which in this case would be made difficult both because of the number of correlated variables and the relative lack of information about which variables are most likely to be useful to the model. The latter, in contrast, involves creating a set of new variables transformed from the old variables in such a way that the variability in the old data set is captured, but the new variables are independent and non-correlated to each other (p. 761, Field et al. 2012). This can be achieved by a technique known as principle component analysis (PCA), and has a beneficial side effect in that it often allows the total number of variables to be reduced by dropping those components contribute least to the overall variability in the data-set (p. 762, Field et al. 2012). It's major downside is the new variables are not interpretable in the way that the old variable set is, as each new variable represents a combination of multiple of the old variables. PCA preprocessing was performed before logistic regression model training, dropping the least important components such that 90% of the original variance was maintained. All data was centred and scaled before the PCA was performed.

Machine learning algorithms do not always gracefully handle missing data (p. 70, Witten 2017) and it was hence necessary to correct for this. Only a single participant had missing data for the variables required for the first strategy and this participant was removed from the data set. Much more missing data was present within the variables required for the second strategy, largely due to missing axial images, mostly at the L1/2 and L2/3 levels. This problem was further complicated by the collection of images at those levels being done at the discretion of the performing radiographer and hence their presence or absence being indicative of the radiographers judgement as to whether those levels were potentially of pathological importance — information it is necessary to mask from the machine learning algorithms (p. 62, Witten 2017).

Initial consideration was given to simply excluding the variables collected at the L1/2 and L2/3 levels, but this strategy would make the remaining data poorly represent those participants with narrowing at those levels and hence was rejected. Alternatively, missing data could be replaced by a stock "normal" value, assuming if the level was not imaged that it was not pathological, but judging what this value should be is fraught with difficulty and also risks the algorithms learning to use that value as a marker of the radiographers decision. Similar issues plague use of an average measurement taken from the available measurements within each participant – this ignores the normal differences in the spinal canal between levels and would likely underestimate the size of the relatively "non-pathological" levels that were not imaged compared to the more likely to be imaged "pathological" levels.

Instead a technique known as k-nearest neighbours imputation was used (Venables et al. 2002). In this technique missing values are replaced by averages taken from the same measurements in a cluster of other participants who are close to the participant in question in terms of their other available variables (p. 42, Kuhn et al. 2013). This strategy avoids re-

moving variables with missing data, replaces missing values in a way that accounts for the level at which they are taken, and avoids using values that are similar enough to be potentially used as a marker of the missing nature of the data at that level. While this technique is also likely to underestimate the true unmeasured value of the measurement in most participants, again due to the greater likelihood of imaging being available at those levels showing narrowing, enough levels with relatively minor degrees of narrowing should be available to mitigate this, particularly at the upper spinal levels where few measurements met previously discussed criteria for stenosis (see previous results chapters).

8.3.3 Strategy 1 results

Strategy 1, as previously described, involves training an ML models using only summaries of the most severe and mean canal diameter per participant. For reference, the AUC for these measurements for separating NC and LBP participants is shown in table 8.2.

The CART algorithm was trained with tuning of three hyper-parameters: the cost complexity factor (determines the aggressiveness of tree pruning), the maximum depth of the tree and the minimum number of data points within a nodes required for the node to be split further. The results of this tuning process are shown in figure 8.6. In general algorithm performance increased up to a maximum performance as the minimum node size approached 30 and in maximum tree depths above 8. Cost complexity had little effect on mean AUC within the central part of its range. The tree selected from the training runs using the "one-standard error" heuristic described above had a cost complexity of 1×10^{-10}, a tree depth of 4 and a minimum node size of 30 and is shown in figure 8.5. Interestingly, the tree largely relies upon Bartynski grading of nerve root impingement in the lateral recess and mean neural exit foramen diameter. Central canal measurements are only used in one node which splits on the mean Schizas grade. The DS-CSA is not used within the decision tree. The output from the decision tree in terms of predicted probability of belonging to the NC participant group across all training folds is shown in figure 8.7. The selected model had a mean AUC of 0.73 (95% CI: 0.70 – 0.76).

The random forest algorithm was trained with tuning of two parameters: the number of predictors randomly sampled at each split when creating the trees and the minimum number of data points within a node required for that node to be split further. The number of trees within the forest was held constant at 1000 across all trained models. The results of the tuning process are shown in figure 8.8. As expected the algorithm was robust to hyper-parameter setting (note the small range of the y axis) with a trend towards increased AUC with increased minimum node size and reduced number of sampled predictors per split. The final selected model had a minimum node size for further splitting of 18 with 2 predictors randomly sampled for each split. This model had an mean AUC across training folds of 0.78 (95% CI: 0.75 – 0.81) and its probability outputs are shown in figure 8.7. The importance of the different variables to the trees predictions was estimated and is shown in figure 8.9. Compared to the more simple CART algorithm the random forest model relied much more heavily on the Schizas grade, while retaining the importance of the Bartynski grade of lateral nerve root impingement.

The logistic regression model was trained while tuning a single hyper-parameter: the total amount of regularisation (penalty) in the model. The results of the tuning process are shown in figure 8.10. The effect of the penalty parameter was non-linear but overall the logistic regression was fairly robust to the hyper-parameter setting. The model with a penalty of 1×10^{-4} was chosen which had a mean AUC of 0.76 (95% CI: 0.73 – 0.80). Again, its probability output across training folds is demonstrated in figure 8.7.

8.3.4 Strategy 2 results

Strategy 2, as previously described, involves training ML models for classification of participants into NC and LBP groups using the raw measurements taken from the MRI studies, without the use of summary variables such as mean or minimum measurements.

A neural network with 5 units in the hidden layer was trained with tuning of two hyper-parameters: the number of training epochs (i.e. the number of cycles through the training data each neural network was exposed to during training) and the proportion of parameters randomly dropped out of the model during training, a process which helps prevent overfitting. The results of the tuning processes are shown in figure 8.11. The selected model used 100 training epochs and a dropout of 0.2. On the training data it averaged an AUC of 0.81 (95% CI: 0.78 – 0.84). Of note, the probability output from the selected neural network (figure 8.7), unlike the other algorithms discussed thus far, did not center on 0.5 but was shifted to the right, meaning the default 0.5 probability cut of for predicting participants as being in the LBP or NC group was inappropriate. To correct for this, Youden's J statistic was calculated for all points on a ROC curve drawn from the training fold performance and the probability cut off which maximised the J statistic (a threshold of 0.842) was selected for use in the test stage (figure 8.12).

Random forest models were also trained for strategy 2, tuning the same hyper-parameters used in strategy 1 (figure 8.13). With the raw data the random forest algorithm again showed a trend towards increasing mean AUC with reduction in the number of randomly sampled predictors per split. There was however no clear trend in the effect of the minimum node size. The selected model had a minimum node size of 2 and 1 randomly sampled predictors per split. This model had an average AUC of 0.78 (95% CI: 0.75 – 0.81). Figure 8.14 shows estimated input variable importance for the selected model. Again, qualitat-

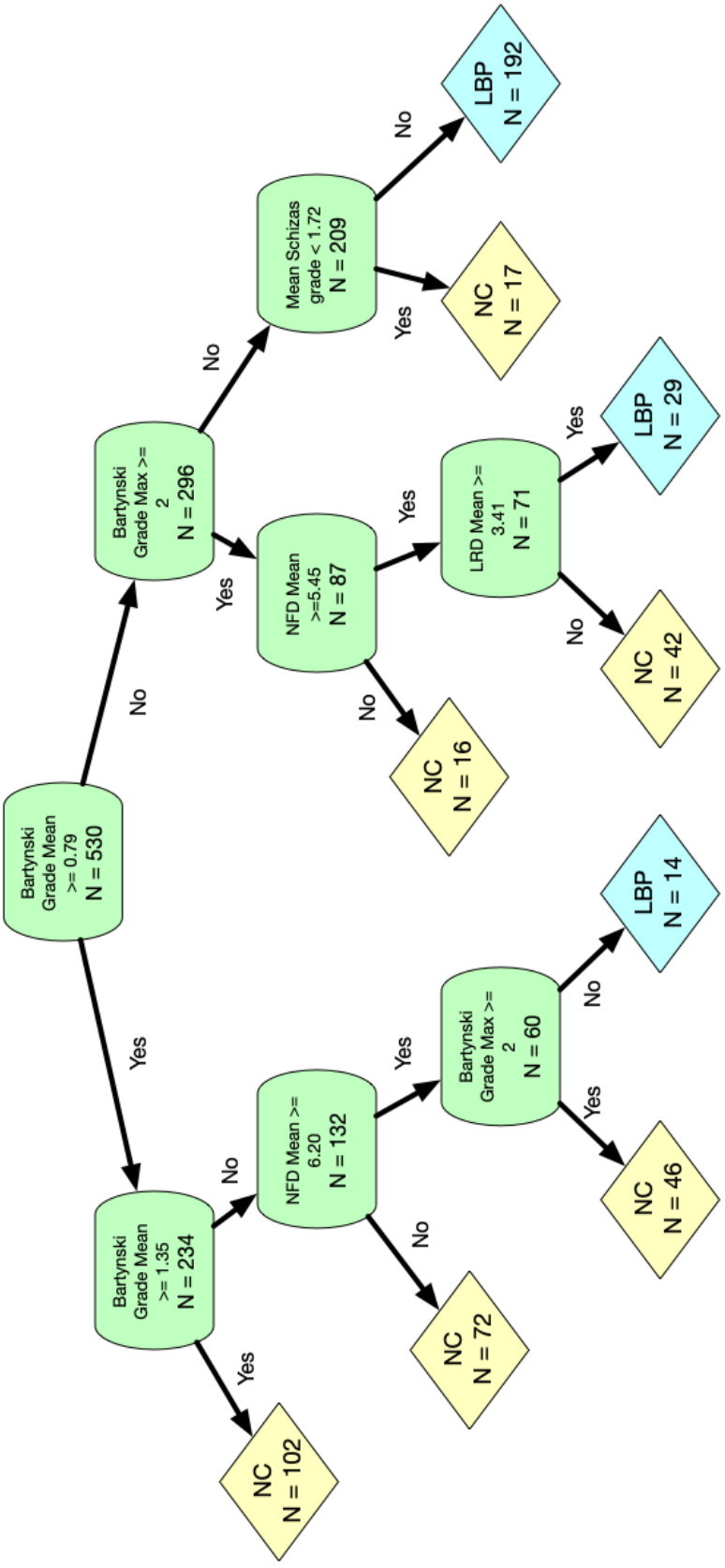


Figure 8.5: The structure of the decision tree selected from those trained using the "one-standard error" heuristic. Notes: N in the root node is larger than the total participant number due to the ROSE re-sampling process; single classification outputs have been shown for simplicity while in reality the tree outputs a predicted probability for group membership for each terminal branch.

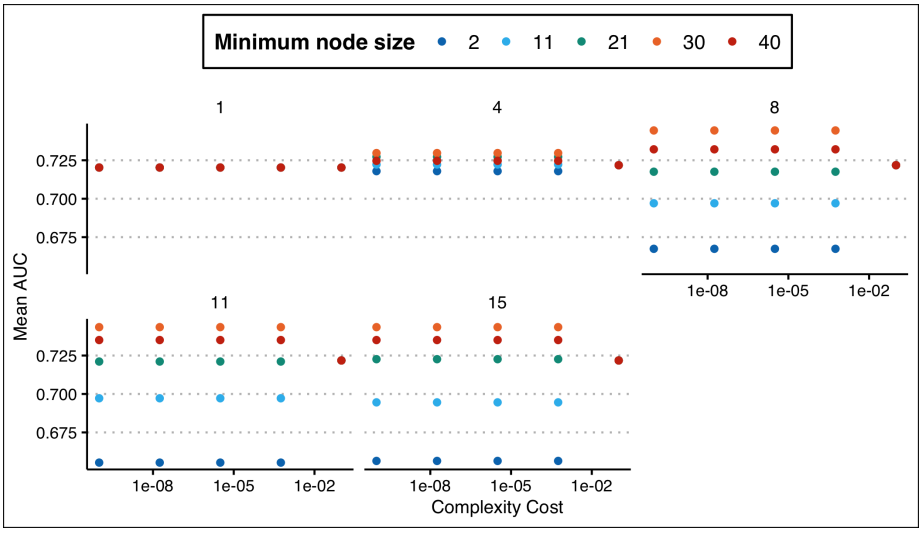


Figure 8.6: Results of hyper-parameter tuning for the CART algorithm. Mean AUC across training folds is plotted against complexity cost, with separate plots for maximum tree depth and colour representing minimum node size.

ive grading systems were judged more important, the Schizas grade at L3/4, and grading of nerve root impingement in the lateral recess and lower neural exit foramina playing the most important roles in model predictions. In a similar way to the probability output seen for the selected neural network model, the predicted probability of NC group membership output from the model were again not centred on 0.5, this time shifted slightly to the left (figure 8.7). An optimum output probability threshold of 0.448 was found using Youden's J statistic.

Given the relatively poorer performance of the CART algorithm in strategy 1, and instability of logistic regression fit on the raw data (showing very marked variation with hyper-parameter tuning) the performance of these algorithms were not explored further in strategy 2.

8.3.5 Final model selection and test performance

A summary of the selected model performance for all algorithms across both strategies is shown in figure 8.15. Based on this, the neural network and strategy 2 random forest based model's were chosen for final evaluation on the test set.

ROC curves for the performance on the test set is presented in figure 8.16. The AUC for the final neural network model's performance on the test set was 0.71, and the AUC for the final random forest model was 0.75, both showing a significant fall in AUC compared to there predicted accuracy in the training stage; almost certainly representing overfitting, more pronounced for the neural network, despite the efforts taken to prevent this. At the predicted probability thresholds selected from the training data the neural network model showed a sensitivity of 77% and a specificity of 60% and the random forest showed a sensitivity of 60% and a specificity of 80%. Sadly, their performance did not improve upon that of the summary measurements discussed in earlier chapters.

8.4 Modeling nerve root compression

8.4.1 Introduction

As discussed in the Chapter 1, the generally presumed mechanism for generation of NC symptoms is nerve root compression. In addition, some have theorised that multi-level low-level compression plays a role in symptom generation through inhibition of venous return to both draining veins at the neural exit foramen and conus R. W. Porter et al. 1992.

If both NC symptoms are truly related to the presence of nerve root compression in the lumbar spine (with NC symptoms occurring with increasing probability as the number of nerve roots either compressed or multiply compressed for a single individual increases) *and* nerve root compression can be reliably predicted at a given site on an MRI scan, then we can hypothesise that an algorithm to detect and count the number of nerve root compressed or multiply compressed within the lumbar spine based on a set of measurements from a lumbar spine MRI would be of utility in separating NC from LBP patients and would do so with higher accuracy than the use of the single measurement types assessed in previous chapters. In, addition, we can hypothesise that the number of nerve roots judged to be compressed or multiply compressed would relate to the degree

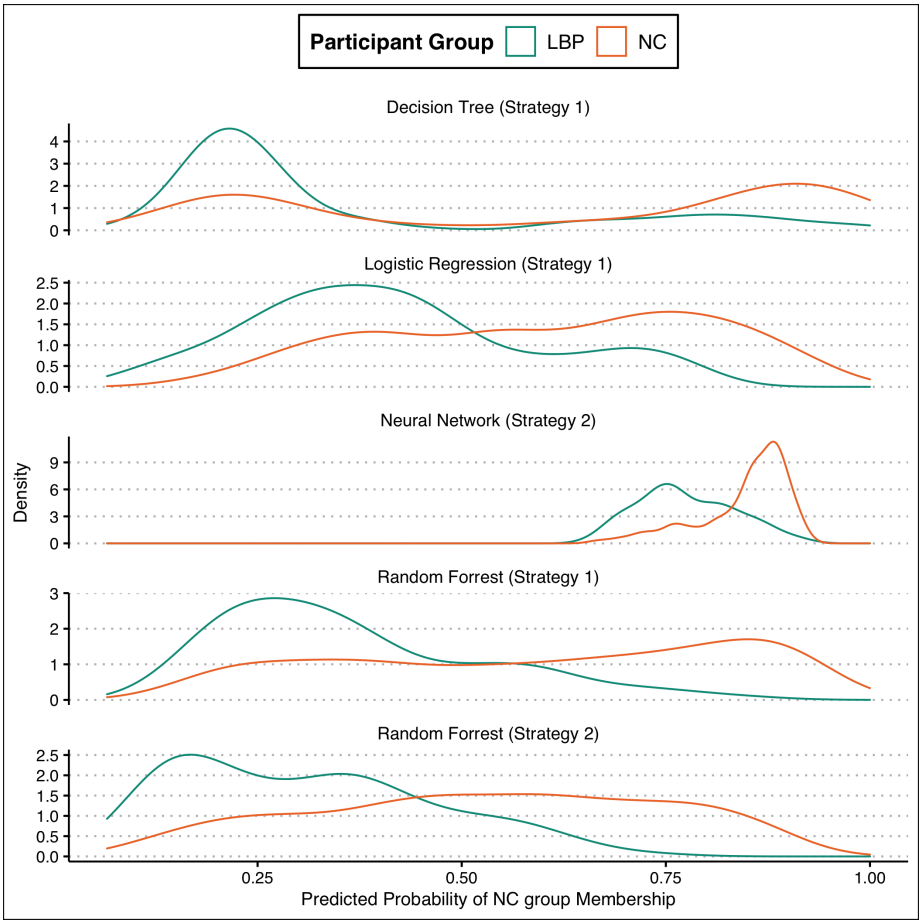


Figure 8.7: Graphs displaying the predicted probability of participant group membership across training folds for the models selected after hyper-parameter tuning.

of symptom severity within patients with NC symptoms.

An algorithm to assess the degree of nerve root compression within the lumbar spine could be imagined to follow the following steps:

- For each nerve root extending through the lumbar canal:
  - For each lumbar level, moving from L1 to L5:
    * If the nerve root is in the central canal: Check central measurement for compression
    * If the nerve root is in the lateral recess: Check lateral measurement for compression
    * If the nerve root is in the neural exit foramen: Check foraminal measurement for compression
    * If the nerve root has left the canal: Continue to next nerve root
  - Based on the above judge whether the nerve root is compressed at a single site or multiple sites
- Based on the above count the number of nerve roots judged to be compressed and/or multiply compressed.

When trying to implement such an algorithm, we would however not know at what threshold level of a measurement to consider indicative of nerve root compression or how to combine information from the qualitative and quantitative information sources available to us. However, if the algorithm does indeed work and is predictive of the presence of NC with some combination of thresholds and a given rule on how to combine the measurements, this combination can be found by brute force – trying all possible combinations of thresholds and selecting the combination that gives the algorithm the the most predictive power. Care must be taken that such a process does not simple choose a combination of thresholds that happens to work because it fits the quirks of the data-set rather than being representative of the underlying disease process (in a similar way to the overfitting problem seen in machine learning section above). We can have more confidence that this has not occurred through following a similar procedure to that used in the following section, by splitting the available data into a training and test-set, selecting the threshold combination based upon its performance on the training-set and then seeing if it retains the same performance in a withheld test-set.

Of note, while such an algorithm could potentially be arrived at in alternate form by the ma-

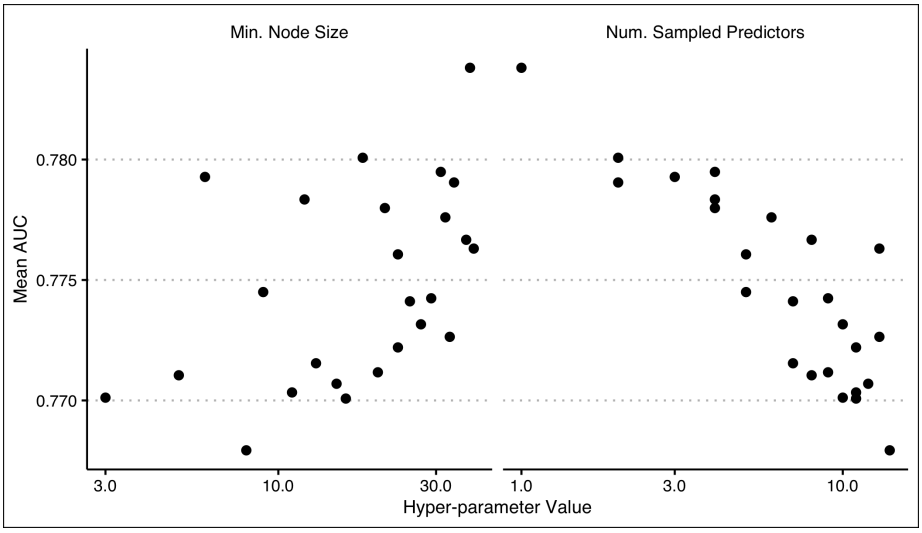


Figure 8.8: The results of hyper-parameter tuning for the random forest based models trained on summary measurement data. The graph on the left shows the effect of minimum node size, while the graph on the right shows the effect of the number of sampled predictors. Note the logarithmic X axis and small range of the Y axis.

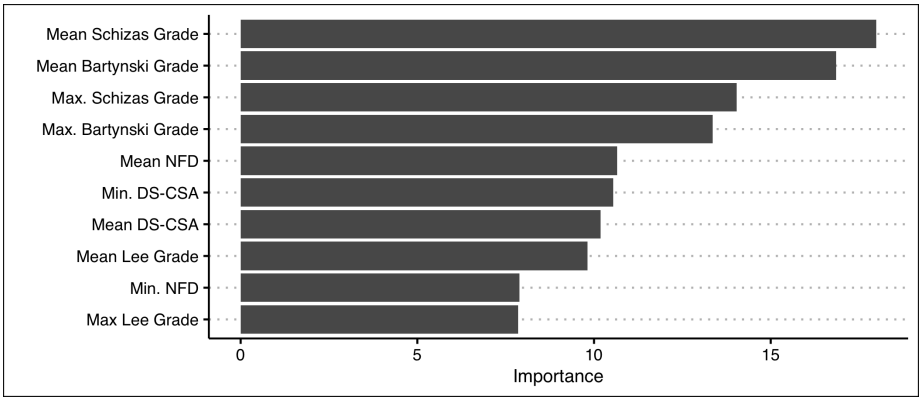


Figure 8.9: An estimate of variable importance to the selected random forest model trained using summary measurement data. Bars for the highest ranking 10 variables are shown.

chine learning algorithms discussed in the previous sections, it would require a significant degree of model complexity. Generation of highly complex algorithms require a large amount of data for training to avoid overfitting. Overfitting was seen within the selected ML models discussed above despite the specific attempts to reduce model complexity. We can therefore assume that either such a model is not predictive of NC/LBP group status (and therefore was not produced by the training process), or was too complex to be generated given the controls on model complexity and size of the training data.

The following sections describe the attempt to produce a software program to implement the algorithm above, find the best threshold combination for its application, and test its accuracy in predicting the presence of NC symptoms and associated symptom severity.

8.4.2 Algorithm implementation

The full code for the implementation of the algorithm above, written in Python 3, is included in Appendix D and illustrated schematically in figure 8.17.

Within the program related data and the functions which act on that data are grouped into classes. A particular instance of a class is called an object (for example, for a class called dog, a particular dog called Rufus might exist and would be considered an object of that class). Within a program multiple objects of a particular class can be created, each storing their own data and functions in a template laid out by the design of that class. This programming methodology is known as object orientated programming (Kindler et al. 2011).

All collected MRI data was stored in a table, with one row per participant, and each column representing a measurement taken from that MRI scan.

A class called participant was created to hold all information relating to a particular real-life participant within the data-set. When the program is first run, one participant object is created for each row in the table of MRI data.

Each participant object once created, stores the measurements taken from that participants MRI in an object of a class called measurements, which exists to make it easier for other parts of the program to access that data. The measurements object extracts from the

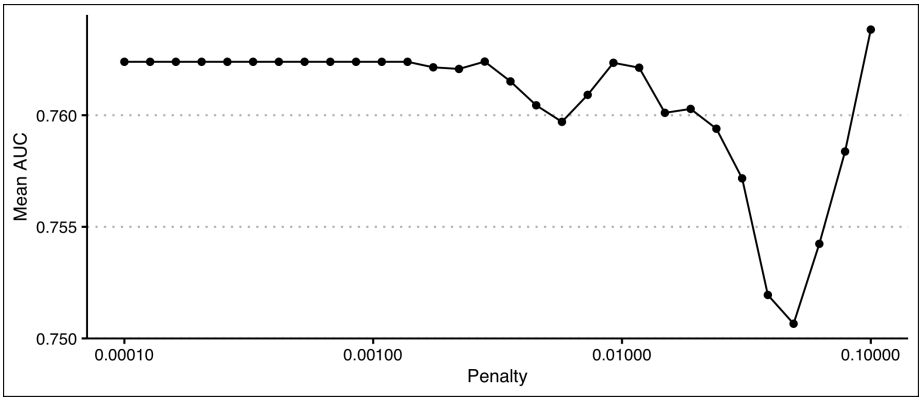


Figure 8.10: The results of hyper-parameter tuning on the logistic regression models trained on summary measurement data.

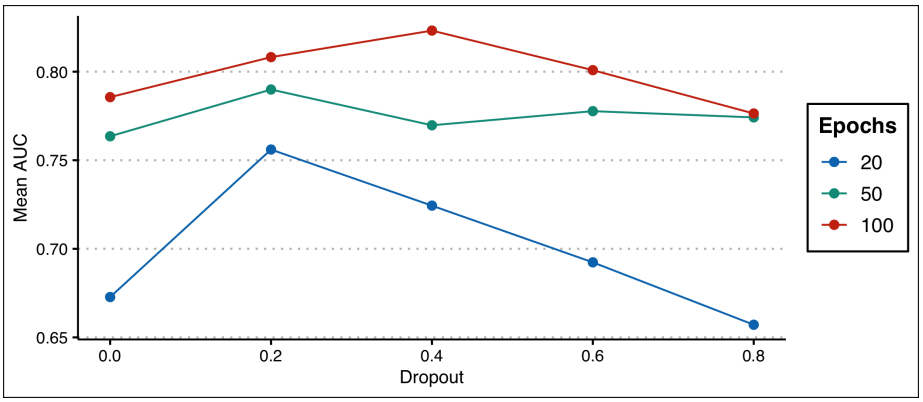


Figure 8.11: The results of hyper-parameter tuning on the performance of the neural network models trained on raw measurement data.

data table the DS-CSA, Schizas grade, LRD, Bartynski grade, NFD and Lee grade for each lumbar level and side for that participant where available — i.e. one quantitative and one qualitative measurement for each central canal, lateral recess and neural exit foramen per spinal level.

The participant object then creates and stores two objects of a class called nerve root for every spinal level between L1 and S2, one for each side of the body. The nerve root class is designed to store information related to the measurements from the MRI study relevant to that nerve root.

The nerve root objects use their spinal level and side to work out the sites in the canal where they can be compressed: 2 levels above their exit from the canal they consider themselves to be in the central canal, 1 level above their exit they consider themselves to be in the lateral recess, and at the level of their exit they consider themselves to be in the neural exit foramen. Once this is done the nerve root objects then get the relevant quantitative measurement of canal size and qualitative impression of nerve root impingement from the measurement object for each site (central canal, lateral recess or neural exit fora-

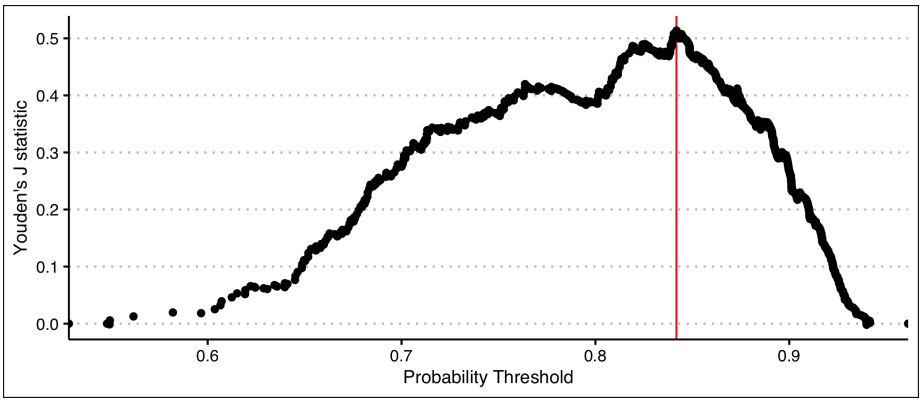


Figure 8.12: A plot of Youden's J statistic against neural network output probability for NC group membership. The statistic was calculated from all points on the ROC curve for the performance of the selected model during training folds. The red line represents the chosen probability threshold for use in the test phase.

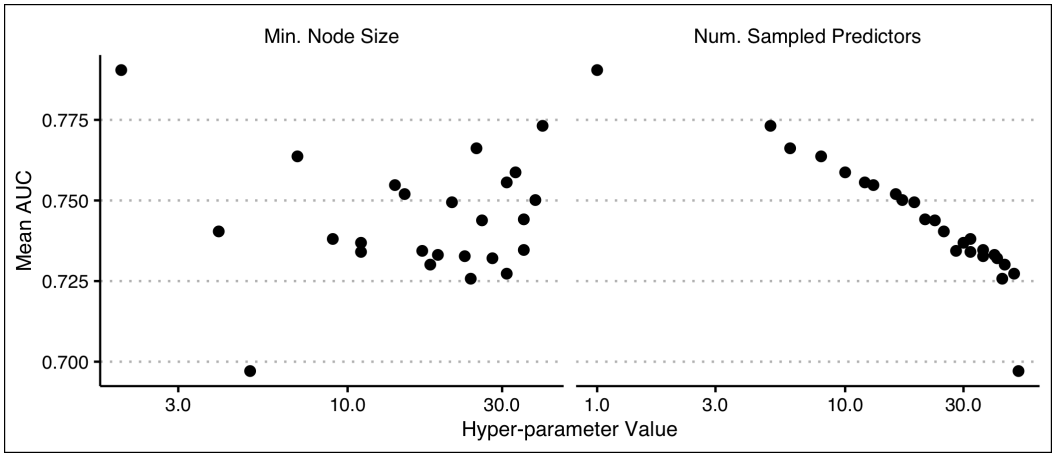


Figure 8.13: The results of hyper-parameter tuning for the random forest based models trained on raw measurement data. The graph on the left shows the effect of minimum node size, while the graph on the right shows the effect of the number of sampled predictors.

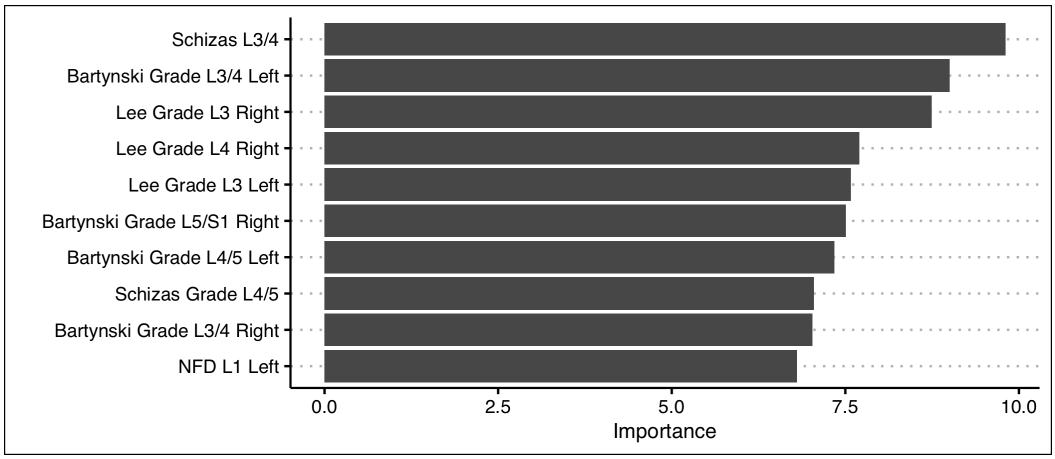


Figure 8.14: An estimate of variable importance to the selected random forest model trained using raw measurement data. Bars for the highest ranking 10 variables are shown.

men) where they could potentially be compressed.

The user specifies a range of possible thresholds and a step size within that range to be tried for each measurement of canal size (e.g. DS-CSA between 50 and 160 mm^2, with steps of 10 mm^2, all Schizas grades between A and D, etc…). The program then iterates through all possible combinations of these thresholds.

Each combination of thresholds is passed to each to each participant object in turn. The participant objects then checks each of its associated nerve root objects against that combination of thresholds and counts the number of nerve roots predicted to be compressed at one site or more (the number of nerve roots compressed (NNC)), and the number of nerve roots predicted to be compressed at more than ones site (the number of nerve roots multiply compressed (NNMC)) for each participant.

When a nerve root object decides if it is compressed at a given site within the canal it makes this decision based on a rule about how it should combine the information from the available quantitative and qualitative measurements. One of four possible rules are used, decided by the user in advance:

- Rule QUANT — if the quantitative measurement at the site of possible nerve root compression is below the provided threshold for that measurement, the nerve root is considered compressed at that site. Qualitative measurement data is ignored.

- Rule QUAL — if the qualitative grading scheme at the site of possible nerve root compression is beyond the provided threshold for that grading scheme, the nerve root is considered compressed at that site. Quantitative measurement data is ignored.

- Rule AND — in order for a nerve root to be considered compressed at a given site, the quantitative and qualitative measurement at that site must meet their respective thresholds.

- Rule OR — in order for a nerve root to be considered compressed at a given site, either the quantitative or qualitative measurement at that site must meet their respective thresholds.

The final output for the program consists of a table of every threshold and combination rule tried

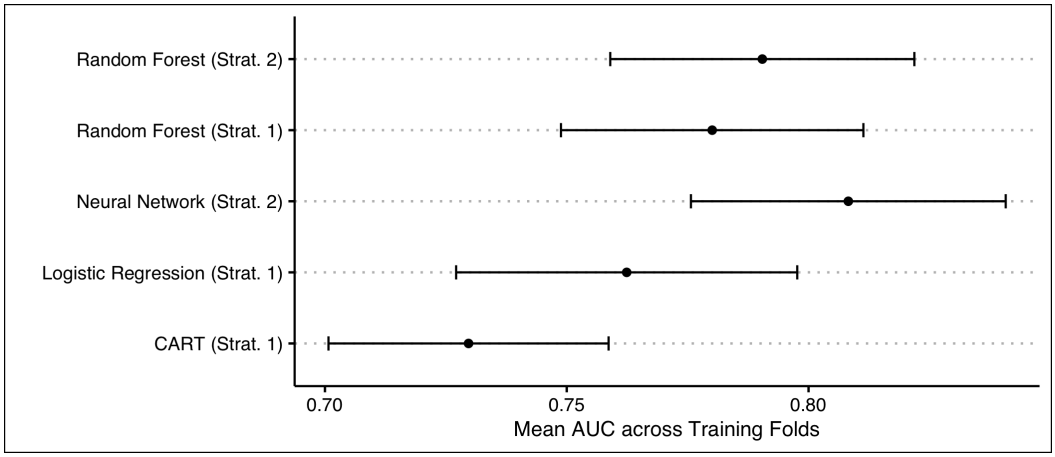


Figure 8.15: A summary of selected model performance in the training stage, with the dot representing the mean AUC and error bars showing the 95% confidence interval.

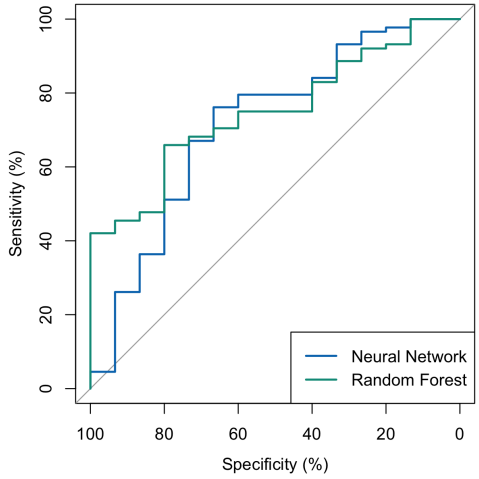


Figure 8.16: ROC curves showing final model performance on the test data.

with the associated predicted NNC and NNMC for each participant.

The program was initially run using the QUANT and QUAL rules with the thresholds ranges below:

- Quantitative thresholds:
  - DS-CSA between 30 and 180 mm^2 with step size of 10 mm^2
  - LRD between 0 and 8 mm with step size of 1 mm
  - NFD between 0 and 8 mm with step size of 1 mm
- Qualitative thresholds:
  - Schizas grades between A and D
  - Bartynski grade between 0 and 3
  - Lee grade between 0 and 3

To ensure the program was working as expected, the mean predicted NNC and NNMC was plotted against input threshold values for each measurement (figure 8.18 and 8.19). As the various thresholds assessed increased in size for quantitative measurements, or decreased in severity for qualitative grading schemes, both the mean NNC and NNMC predicted by the program would be expected to increase as more and more nerve roots fulfilled the criteria for single site compression and multi-site compression. This would continue until the point that all nerve roots that could be considered compressed were considered compressed for each participant. The output from the algorithm was consistent with this prediction suggesting that the program was working as intended. [1] For each threshold of any one measurement tried the graphs demonstrate a large range of predicted NNC and NNMC values — this effect is caused by the interaction with the other thresholds also being tried by the program for any one point. This is demonstrated in figure 8.20, where the points on the graph of the mean predicted NNC per participant has been plotted against the DS-CSA threshold used and each point has been coloured according to the LRD and

[1] In reality the first time the program was run the NFD threshold did not appear to have any effect on the number of nerve roots considered compressed. After some bug hunting this was found to be due to a typographic error in the part of the program that checked the NFD measurements against their associated threshold. The mistake caused the program to never consider a nerve root compressed within the foramina when using the QUANT or AND combination rules.

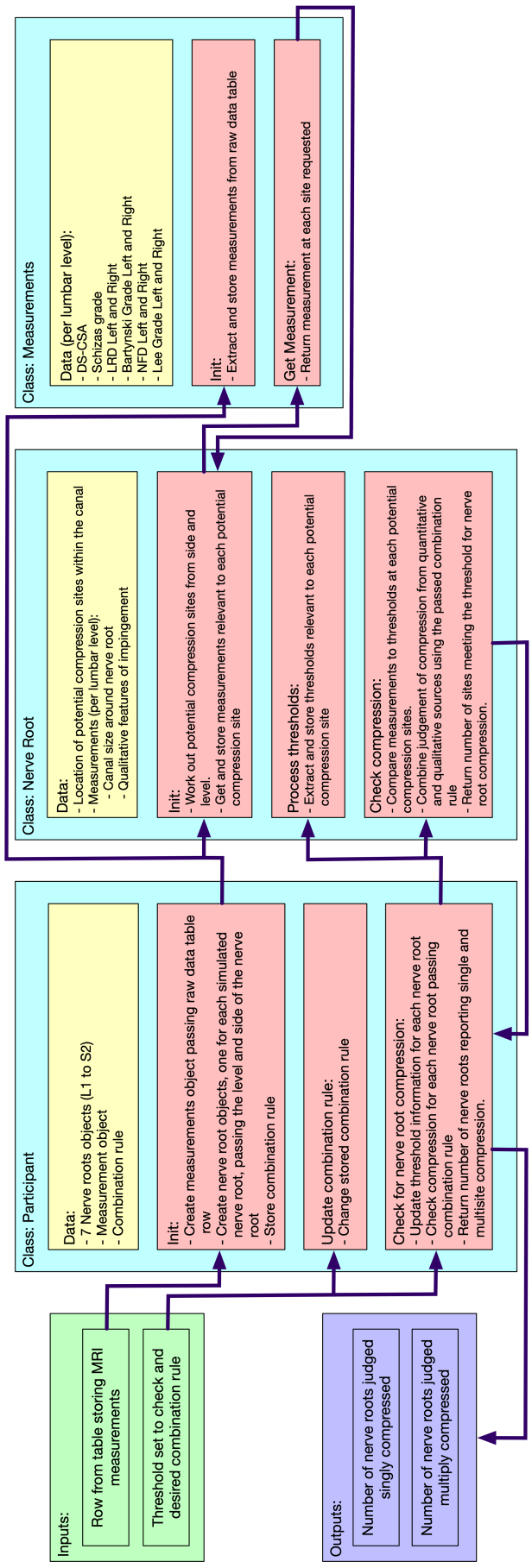


Figure 8.17: The basic structure of the program written in Python 3 to predict nerve root compression based upon the data collected from lumbar spine MRIs. Each blue box represents a class which collects together data (yellow boxes) and a set of functions (red boxes). Inputs to the algorithm are shown in the green box, and outputs in the purple box. Arrows show information flow between the various functions. "Init" are run when an object of a particular class is created.

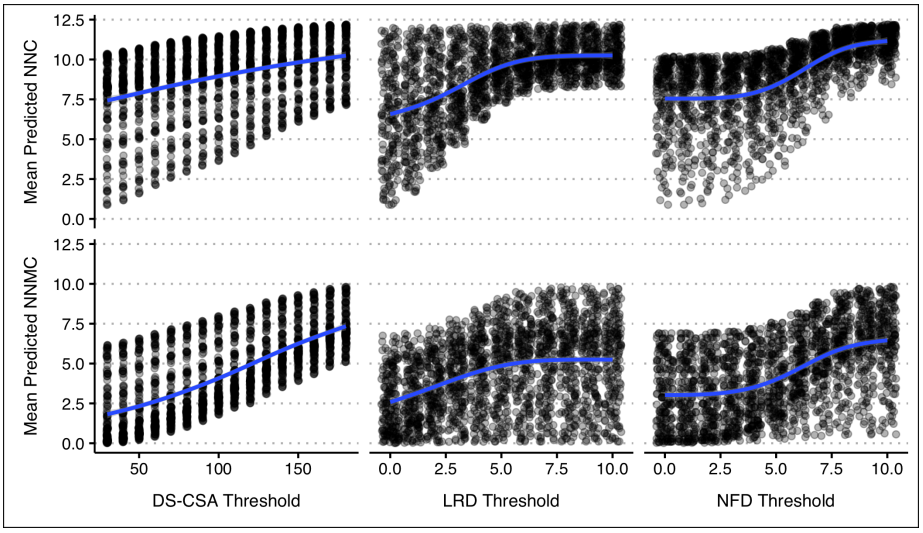


Figure 8.18: The mean NNC and NNMC per participant is plotted against the quantitative measurement thresholds tested by the program. The blue lines are drawn to highlight the general trend in the data. The data was produced using the QUANT combination rule to remove the effect of the qualitative grading measurements.

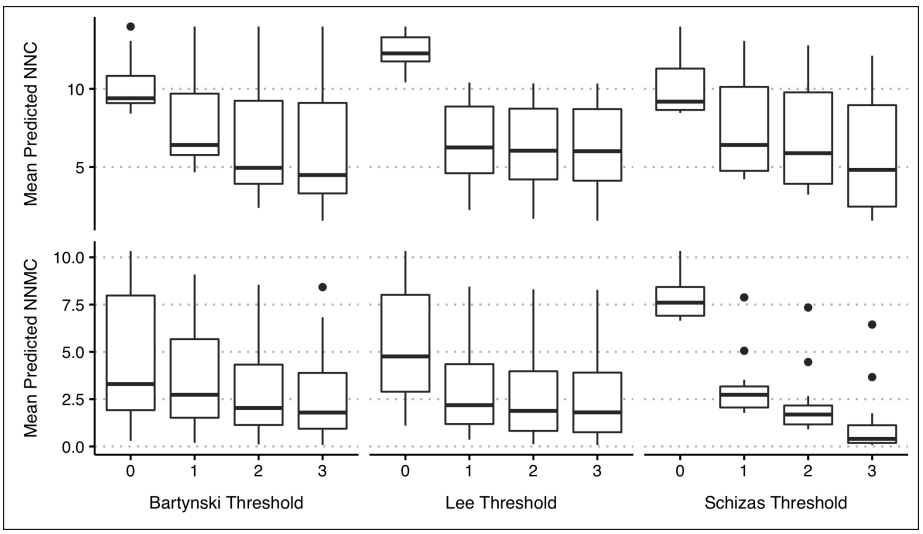


Figure 8.19: The mean NNC and NNMC per participant is plotted against the qualitative grading scheme thresholds tested by the program. Separate graphs are shown for predicted single and multi-level nerve root compression. The data was produced using the QUAL combination rule to remove the effect of the quantitative measurements. All grading systems were converted to 4 point integer scales between 0 and 3.

NFD thresholds used for the same point.

8.4.3 Choosing the best threshold combination

After the program has been run, two ROC curves can then be build for each threshold combination: One assess the ability of the number of the predicted NNC per participant to separate LBP and NC participants and the second assesses the ability of the predicted NNMC per participant to separate LBP and NC participants.

In order to select the best combination of possible thresholds for use with the program, the AUC for the ROC curves for each combination of thresholds was plotted against the threshold values used for each measurement.

First, the output data from the run of the program using the QUANT combination rule was considered (figure 8.21). Using this rule removes the effect of the qualitative thresholds and measurements, meaning the effect of the DS-CSA, LRD and NFD threshold can be seen in relative isolation. The AUC could be seen on average to reach a peak when using a DS-CSA between 60 and 100[2] and a NFD of 6 or 7 mm with little difference between the plots for the predicted NNC or NNMC per participant. Interestingly, the AUC generally increased with reducing LRD threshold when using the predicted NNC but that trend was reversed when using the NNMC.

Second, the output data from the run of the program using the QUAL combination rule was considered (figure 8.22) using this rule to remove the effect of the quantitative thresholds and measurements. When a Schizas grade threshold of A was

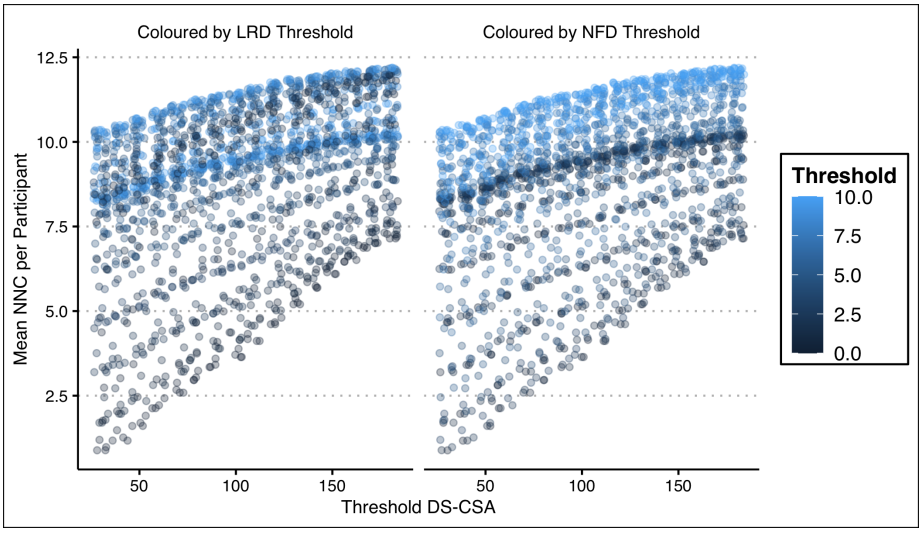


Figure 8.20: The mean predicted NNC per participant is plotted against the DS-CSA thresholds tested by the program. Points are coloured either by the LRD (left) or NFD (right) measurement threshold used by the program for the same data point.

used, i.e. where every nerve root was always considered compressed in the central canal very poor AUCs were seen, as would be expected. A large jump in AUC was seen at thresholds of grades B or above. Similar but smaller effects were seen for the Bartynski threshold and Lee thresholds.

All possible qualitative grade thresholds, and a smaller range of quantitative thresholds, chosen around the peaks in AUC shown in figure 8.21 were then explored using the AND or OR combination rules. The reduction in range of quantitative thresholds to just those likely to contain the maximum AUC achievable was needed to reduce the amount of required computing time. [2]

The performance across all tried combinations of thresholds and combination rules, both when using the predicted NNC and NNMC, is shown in figure 8.23. Use of the NNMC did not improve on AUCs compared to the NNC and the best performing overall combination rule was AND when using the predicted NNC (The OR rule had outliers with higher AUC than those of the AND rule but the median performance across all threshold values was much poorer).

In order to select the best possible combination of threshold measurements to use with the AND combination rule, plots comparing AUC to the input threshold measurements were again created (figures 8.24, 8.25 and 8.26). Median AUC across all thresholds reached a peak with a DS-CSA threshold of 60 mm^2, LRD of 4 mm, NFD of 5 mm, Bartynski grade of 2, Lee grade of 0 and Schizas threshold of 2. The median AUC was used to select the thresholds because it was thought more likely to represent the true effect of the changing threshold, rather than the upper bound of AUC which was thought to have higher risk of being a quirk of the data being evaluated.

The use of a Lee grade threshold of 0 means that all nerve nerve roots would be considered qualitatively compressed within the neural exit foramina by the program. This is prevented by the AND combination rule which also requires the threshold NFD measurement to be met for a nerve root to be considered compressed. Hence with this Lee grade threshold only the neural exit foramen depth has an effect on whether a nerve root is judged to be compressed at foraminal sites. The fact this maximises the AUC is perhaps unsurprising given the poor classification performance of the Lee grade compared to NFD seen when using single measurements alone (see table 8.2). This sort of interaction between the various thresholds and the effectiveness of the given measurements for separating LBP and NC participants is likely behind other effects such as the reversal of the trend of the AUC with changing LRD threshold comparing NNC and NNMC for the QUANT rule data discussed earlier.

This combination of thresholds with the AND combination rule resulted in an AUC of 0.78 (95% CI: 0.72 – 0.84) on the training data, with an optimal cut-off of 2 or more nerve roots predicted to be compressed to predict a participant having NC and giving a sensitivity of 80.0% and specificity of 62.0%. When applied to with-held test data the AUC was also 0.78 (95% CI: 0.68 – 0.88) — figure 8.27. This AUC was higher than the best performing single measurement for identifying NC participants — the mean Bartynski grade with AUC of 0.76 — and the AUC for the minimum DS-CSA per participant (AUC of 0.73), but these differences did not reach significance. There was no significant correlation between the number of nerve roots predicted to be compressed with this combination of thresholds and the patients symptom severity.

[2] A total of 1936 threshold combinations were tried for the QUANT rule, 63 for the QUAL rule (where only 4 possible grades for each of the three grading systems had to be tested) and 20160 combinations for each of the AND and OR rules, even with the reduced range of quantitative thresholds used. For each combination of thresholds the program must predict the compression of 5838 nerve roots (14 nerve roots per 417 participants). Without parallelisation, the program took more than 3 hours to complete on my computer for each of the latter two combination rules, though I'm certain a better programmer could optimise the program extensively and this could be also be improved by switching to a compiled language.

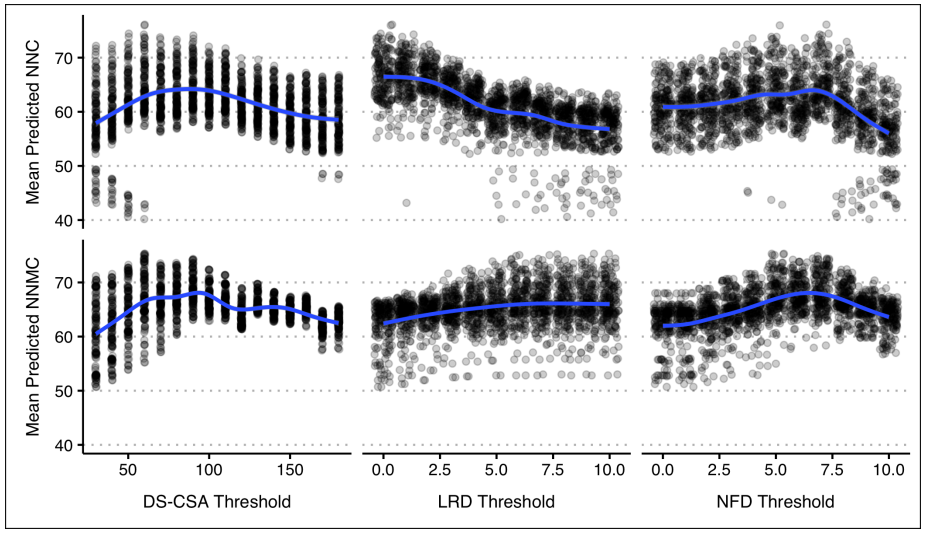


Figure 8.21: The effect of quantitative thresholds on the classification accuracy for NC/LBP group status. Only data from the QUANT combination rule is included, meaning qualitative grading scheme thresholds were ignored when creating this data.

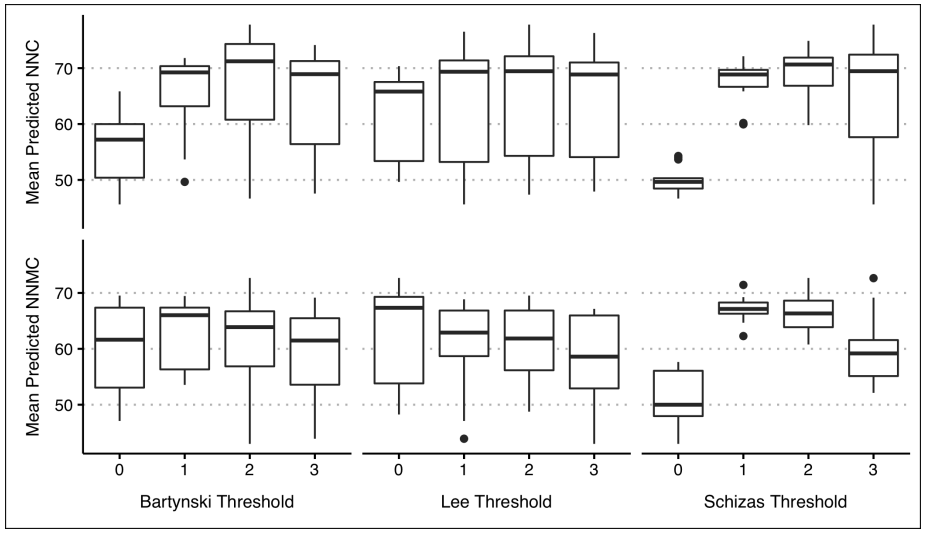


Figure 8.22: The effect of qualitative grading scheme thresholds on the classification accuracy for NC/LBP group status. Only data from the QUAL combination rule is included, meaning quantitative thresholds were ignored when creating this data. The Schizas threshold is shown as a numeric scale with 0 representing grade A, 1 representing grade B, etc.

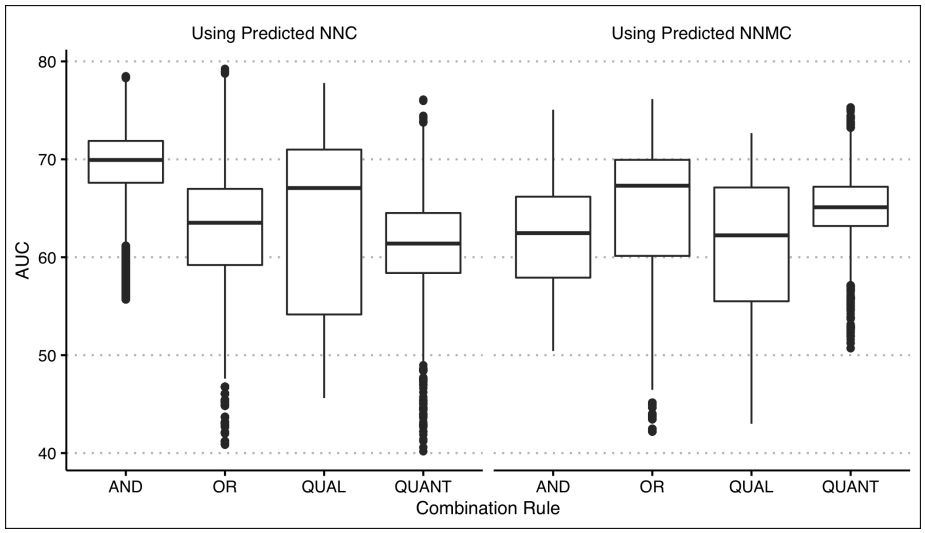


Figure 8.23: The effect of the combination rule used on the classification accuracy of the NNC and NNMC for NC/LBP status.

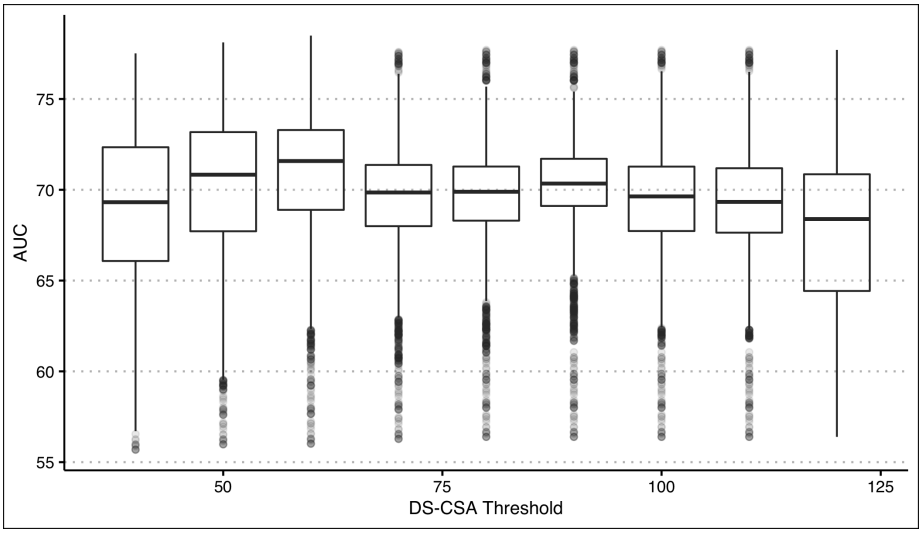


Figure 8.24: The effect of the DS-CSA threshold on the classification accuracy for NC/LBP status of the number of nerve roots predicted to be compressed at at least one site when using the AND combination rule.

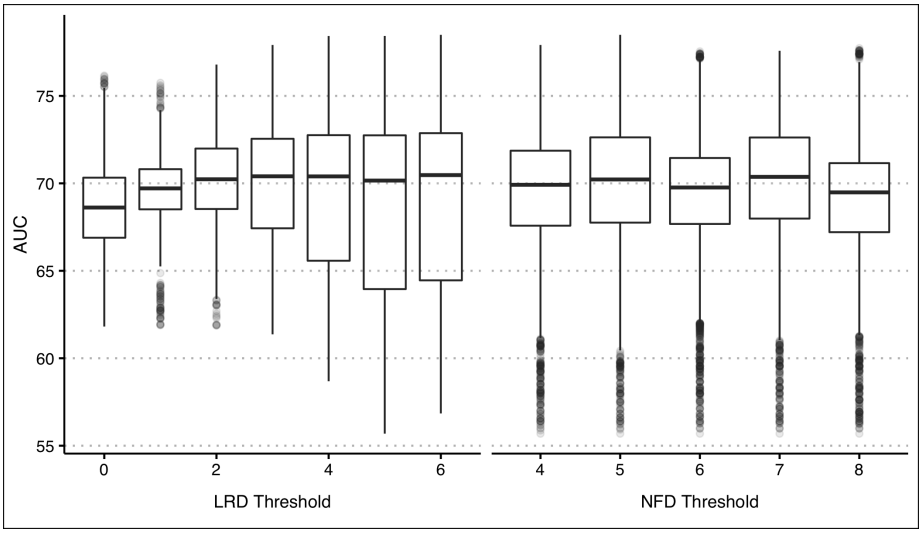


Figure 8.25: The effect of the LRD and NFD threshold on the classification accuracy for NC/LBP status of the number of nerve roots predicted to be compressed at at least one site when using the AND combination rule.

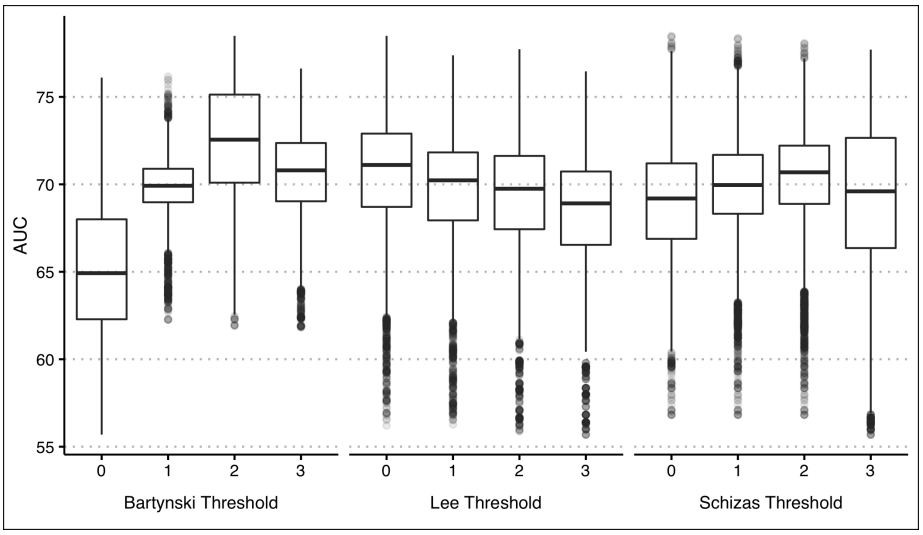


Figure 8.26: The effect of the qualitative grading system thresholds on the classification accuracy for NC/LBP status of the number of nerve roots predicted to be compressed at at least one site when using the AND combination rule.

8.5 Chapter summary

This chapter presented the results of various attempts to combine the numerous measurements taken from each MRI spine to improve upon the ability of the any one measurement type to separate NC and LBP patients compared to the performance seen in earlier chapters.

First, the predicted stenosis ratio, based upon the measurement first suggested by Laurencin et al. (1999), was introduced. This measurement used the APD as a marker of canal size unaffected by degenerative change, and used it to transform the DS-CSA measurements in a way which hopefully made the measurement more representative of the degree of degenerative narrowing and less dependent upon the normal variation in the non-degenerative size of the canal. The resultant transformed measurement more closely correlated with qualitative features of central canal stenosis than the unmodified DS-CSA, but the transformation did not improve its ability to separate NC from LBP participants or its correlation with clinical severity scores.

Second, 4 different machine learning algorithms were trained to separate NC from LBP participants, either on the raw measurements taken from the MRIs of the lumbar spine, or on the mean and most severe measurements per participant. In general the models trained on the raw measurements performed slightly better in the training phase and the neural network and random forest based models trained on this data were then tested on test data, withheld from model training. Despite good classification performance in the training stage, a fall in performance was seen on the test data, likely representing overfitting, and the overall classification performance did not exceed that of the best performing single measurements.

Finally, a software program which attempted to simulate nerve root compression across a large combination of different potential definitions of stenosis in the central canal, lateral recess and neural exit foramen was introduced. Given a set of threshold combinations to check and a rule on how to combine measurements taken at the same site, the program output the number of nerve roots it thought was compressed within the canal of each participant. The ability of these numbers to separate NC from LBP participants was found for each threshold combination, and the best performing combination was found manually by inspection of graphs of the various threshold measurements plotted against the classification performance. The final classification ability of the program with the selected threshold combination was then tested on a test set of withheld data and the overall classification accuracy with the selected threshold set was higher than any previously tried method in this thesis. Despite this, the difference with the best performing single measurements did not reach statistical significance. Of note, only counting nerve roots thought compressed at two or more sites did not improve the accuracy of the program, despite the postulated role of multi-site compression in NC symptom generation.

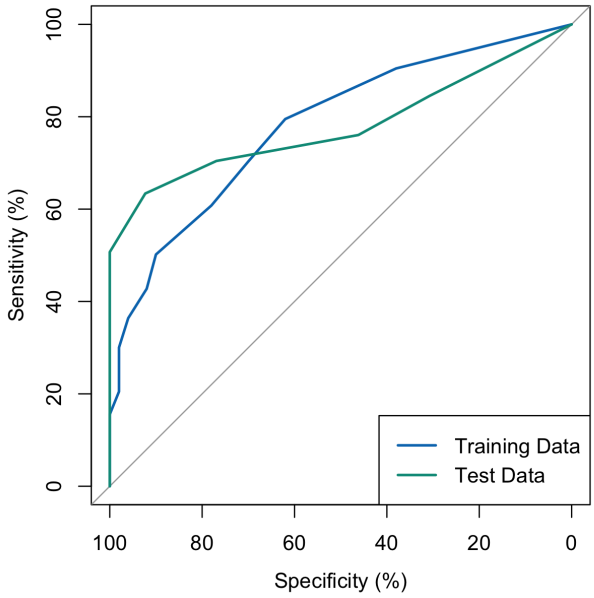


Figure 8.27: The ROC curves for the predicted number of compressed nerve roots produced by the program using the selected threshold set and the AND combination rule.

Chapter 9

Conclusions

9.1 Discussion: thesis aims and research questions

The overall aim of this thesis was to work towards a clinically meaningful definition of anatomical lumbar spinal stenosis, which is currently conspicuously absent from both clinical guidelines (Kreiner et al. 2013) and from the literature in general (Andreisek et al. 2011)

In § 1.7, three main research questions aimed at working towards this goal were defined: what imaging features are associated with neurogenic claudication, is there a relationship between spinal canal size on MRI and severity of clinical presentation in symptomatic LSS, and what is the best way to measure narrowing of the spinal canal on MRI? The contribution towards answering the first two questions made by the results of the presented observational study are discussed in the following sections. The final question will be answered in the final conclusion of this chapter.

9.1.1 What findings are associated with neurogenic claudication?

One of the major challenges faced by research involving imaging based tests for the diagnosis of symptomatic lumbar stenosis is the lack of an accepted reference standard which has clinical relevance that goes beyond simply confirming the accuracy of depiction of anatomy, and instead allows a judgement about the patients true underlying disease status. In § 1.7 the different approaches found in the literature aiming to work around this limitation were discussed alongside their various limitations. Furthermore, it was noted that studies assessing imaging in symptomatic LSS patients in general struggle with implicit bias introduced by the fact the very definition of symptomatic LSS (and hence the patient population studied) includes the requirement for an imaging based judgement about the presence of canal narrowing, however defined. The need for a new approach to finding an definition for anatomical LSS which avoided this problem was expressed, and it was proposed to start from first principles by looking for MRI findings (canal measurement based or otherwise) which separated patients with the key clinical feature of lumbar spinal stenosis, neurogenic claudication, from other individuals, while avoiding patient groups selected based upon radiological criteria or other reasoning about the underlying disease process. The hope was that the measurement that best separated symptomatic from non-symptomatic individuals, without the sources of bias noted above, would potentially have pathological significance and act as a starting point for a definition of anatomical LSS. This definition would then need validation through further studies to assess its prognostic and treatment implications.

In § 2.4, a systematic review of papers comparing imaging findings in a case population with neurogenic claudication to a variously defined control group was presented. 14 papers were included, and generally found evidence of smaller DS-CSAs, effacement of CSF space around the cauda equina and increased ligamentum flavum size. All but one included paper fell into the bias trap described above by studying symptomatic LSS populations and either implicitly or explicitly separating patient groups by imaging based criteria. Only a single paper avoided this, Hamanishi et al. (1994), who showed smaller DS-CSA measurements in NC patients and higher levels of multi-level stenosis compared to the control group. This analysis however was on a vertebral level by level basis, limiting its interpretability and the paper did not attempt to find a threshold DS-CSA measurement for separating the case and control populations. There were numerous other significant flaws in the included papers, including a failure to record whether MRI observers were blinded to the case or control group status of the patients in 11 of the 14 papers including in Hamanishi's.

The observation study presented in Chapter 3 through 5 aimed to address these issues. Its case population was participants recruited as part of the BOOST RCT, which included patients purely based upon a clinical presentation of neurogenic claudication symptoms. Recruitment was from both community and secondary care settings, helpful in including a larger range of potential symptom severity and avoiding accidental selection of participants

with smaller canals through higher rates of GP to secondary care referral or secondary care follow-up in the presence of an abnormal MRI. In addition, MRI observers were blinded to case or control participant status when assessing each MRI study.

The central canal

Three main methods of assessing central canal stenosis were assessed: the APD, the DS-CSA and the Schizas grade.

The APD measurement behaved in a manor consistent with a marker of the developmental size of the canal. It did not correlate with the patients age, and showed no clear trend with qualitative features of degenerative stenosis as represented by the Schizas grade. Almost no included patients met Verbiest's 1954 original definition for developmental stenosis, consistent with the degenerative aetiology of stenosis in our largely elderly patient group. The APD, unsurprisingly therefore, had the poorest power to discriminate NC from LBP participants of the central measurements, with an AUC of 0.62 for the minimum measurement per participant, significantly worse than the performance of the minimum DS-CSA measurement.

The DS-CSA and Schizas grades while performing better than the APD measurements, also showed significant overlap in their distributions between NC and LBP participants with resultant AUC values for both mean and most severe measurement per participant of between 0.69 and 0.73. No significant difference in AUC values was found between these measurements.

Interestingly, the optimum threshold for the minimum DS-CSA per participant, minimising false positives and false negatives equally, was found to be 91.5 mm^2, with a sensitivity of 63.7% and specificity of 73% for the presence of NC symptoms. This threshold lay between the 100 mm^2 and 75 mm^2 values popularised by Schönström as relative and absolute stenosis respectively (Schonstrom et al. 1985). Reducing the threshold to the 75 mm^2, the threshold Schönström thought would reproducibly result in a intra-dural pressure rise (Schönström et al. 1988), increased the specificity to 84.1% (with 52.9% sensitivity).

Similarly, for the most severe Schizas grade per participant, the optimum threshold, minimising false positives and false negatives equally, was found to be Grade B, with 73.0% sensitivity and 62.1% specificity. If instead a grade C was used, the specificity increased to 85.7%, while sensitivity became poor at 48.4%. Grade C and D grades have been suggested as surgically treatable disease by Schizas (as opposed to A and B where he thought conservative management should be tried in preference). This was based on the fact that grade C and D patients seemed more likely to fail conservative management (Schizas et al. 2010) and Grade C and D patients appear to show greater odds of showing good post surgical improvement (Mannion et al. 2017).

One criticism that has been levied at use of quantitative measurements, such as the DS-CSA, for measuring stenosis is that absolute values of such measurements fail to take into account the normal non-degenerate size of the canal for that given patient and spinal level (Andreisek et al. 2011) — the implication being that relative canal narrowing may be more important than absolute canal narrowing.

In § 1.3.4 we noted evidence from the literature suggesting that: taller individuals have larger bony canals, that the evidence for gender based differences in canal sizes was unclear, and that there was reasonable evidence for differences in canal size between different ethnicities. In the currently presented study height was weakly positively correlated with the mean DS-CSA and mean and minimum APD. There were also positive correlations between height and mean lateral recess depth, thought no relationship was found with the neural exit foramen diameter. No significant difference in any central canal or foraminal measurements between gender or ethnicity groups was identified, though minimum and mean lateral recess depth per participant were smaller in men compared to women. We should however note that measurements in our participant groups are likely to be distorted by degenerative change (aside from the APD) and hence may not reflect variation in the non degenerative canal. The study was also not designed to assess for differences in canal measurements between ethnicities and had largely White British participants in both NC and LBP populations, meaning the statistical power for detecting differences in canal size between ethnicities is likely to be low. It should also be noted that absolute differences in canal size between any of these groups, even when statistically significant, were relatively small, both when considering our own study and other published studies.

As noted in § 8.2, the only measurement of lumbar canal narrowing found within the literature which attempted to measure canal narrowing relative to the canals non-degenerate size was Laurencin's stenosis ratio (Laurencin et al. 1999). The ratio was taken as the ratio of the disc level DS-CSA to the pedicular DS-CSA (the latter thought to be relatively invariant with degenerative change). This idea however does not appear to have taken off, with only one further paper to my knowledge making reference to it, finding smaller stenosis ratios in surgically managed patients LSS participants compared to those conservatively managed, but no significant correlation with preoperative ODI or ODI improvement post surgery (Schizas et al. 2010).

It was not possible to measure the stenosis ratio directly on our data set due to absence of axial imaging at the pedicular level on the majority of studies. Instead a technique was described involving fitting a linear model between the APD and the DS-CSA across all participants, and then producing a ratio measurement for each level imaged by dividing the disc level DS-CSA with the DS-CSA predicted from the participants APD at the same level. The idea be-

ing, that the variability of the DS-CSA that could be explained by the APD measurement likely represents non-degenerate variation in canal size and hence by removing it the ratio measurement comes to more closely reflects the degenerative changes and narrowing relative to the non-degenerate size of the canal. In support of this, the "predicted stenosis ratio" was found have a higher strength of correlation with the Schizas grade than the un-modified DS-CSA. Despite this, the predicted stenosis ratio did not show any improvement in differentiating NC from LBP compared to the un-modified DS-CSA. It is possible that this technique failed because the linear model used for producing the predicted DS-CSAs has an over-representation of stenotic levels, given the patient populations included in this study, and that better performance would be observed if it were fitted using a larger, less degenerate spectrum of the population.

Perhaps of more relevance to the quantitative definition of central canal stenosis is the changing volume of neural material in the canal. Nerve roots leave the canal at each lumbar level, meaning the volume of neural tissue within the canal, and hence the DS-CSA required for compression, should progressively reduce moving inferiorly. This effect can be seen in the relationship between DS-CSA and Schizas grade shown in figure 4.6, where progressively smaller median DS-CSAs were required to result in a given Schizas grade of B or above when moving inferiorly. Schömström noted that the apparent volume of neural tissue and other tissue in the canal reduced by a cross-sectional area of around 10 mm^2 between L2 and L4, based upon thresholds of cross-sectional area reduction required for intra-dural pressures to rise in cadaveric specimens. The effect of changing nerve room volume seems comparatively greater in our current study, based on inspection of figure 4.6. A single quantitative threshold defining central canal stenosis will not account for this variable requirement of narrowing to produce compression and no easy way of accounting for this effect with quantitative measurements seems possible without having different definitions of stenosis per spinal level. These problems are in favour of the qualitative grading systems like that of Schizas et al, which are not vulnerable to these problems.

On a related point, quantitative measurements of the canal also face the criticism that they are dependent upon slice angulation relative to the canal. Henderson et al. (2012) found by reconstructing 3D MRI sequences that the DS-CSA could vary by up to 32% within even 10° of induced slice angulation, well within the range of error in slice angulation relative to the dural-sac likely to be induced by normal slice placement by radiographers. In contrast the Schizas grading was found to be highly robust to changes in slice angulation of up to 30°. This must however be balanced against measurement reliability — we found excellent inter-rater and intra-rater reliability for the DS-CSA but a more moderate reliability for the Schizas grade, which obviously depended upon more subjective judgement for assessment. Several weaknesses regarding edge cases which likely contributed to the lower reliability were discussed in § 4.1. Work to provide further definition on how to handle such cases and further observer training would likely improve upon the presented reliability scores.

The lateral recess and neural exit foramen

Studies investigating symptomatic LSS often focus on investigating the pathological significance of central canal narrowing. Of the studies included in the scoping review presented in § 2.1, the few that included a defined radiological definition of stenosis used central measurements of the canal. Similarly, for those studies investigating the relationship between NC symptoms and control groups, no included paper studied the effects of lateral recess or foraminal measurements. Lateral recess stenosis and foraminal stenosis seem often considered a related but separate pathological entity linked to compression of a single nerve root (C. K. Lee et al. 1988), perhaps because in isolated stenosis, they typically only involve a single nerve root and therefore are thought of more in terms of radiculopathy type presentations in the minds of clinicians where clinical syndromes such as NC are more commonly linked in the secondary literature to more generalised cauda equina compression (Katz 2008).

In the presented observational study, close relationships were seen between measures of central canal stenosis and those of the lateral recess and neural exit foramen (§ 5.3 and 6.3). The lateral recesses depth was strongly positively correlated with the DS-CSA, and nerve root entrapment in the central canal (i.e. a Schizas grade greater than B) was strongly predictive of nerve root impingment in the lateral recess (a Bartynski grade of greater than 2). Similarly, the neural exit foramen depth correlated with both the DS-CSA and LRD from the same side, and qualitatively judged central or lateral recess nerve entrapment was also predictive of neural exit foramen entrapment at the same level and side (though these relationships were less strong than between the central canal and lateral recess). These relationships are not unexpected given that the same degenerative processes contribute to narrowing all three structures but highlight that patients with some degree of central canal narrowing are also likely to have narrowing in other parts of the canal that may be contributing to their presentation.

Measurements of the lateral recess depth were significantly smaller, both in terms of the mean and minimum per participant in NC participants compared to LBP participants. AUC values, however, were poor and similar to those for the APD, with values of 0.66 and 0.63 respectively. In contrast the Bartynski grade of nerve root entrapment in the lateral recess performed significantly better, with AUCs of 0.77 and 0.74 for the mean and minimum grade per participant respectively, the former being the best observed performance for a single measurement across those tested in the thesis. The difference when com-

pared with the AUCs for the DS-CSA and Schizas grade measurements did not however reach significance. A threshold maximum Bartynski grade per participant of grade 3 (complete lateral recess obliteration) had a specificity of 84.1% and sensitivity of 57.3%. A mean Bartynsi grade of >= 1.82 had a specificity of 82.5% with a sensitivity of 61.5%.

The discrepancy in performance between the Bartynski grading scheme and the LRD may relate to several problems identified with the LRD measurement and described in § 5.1. LRD reliability was below that of the Bartynski grade, with a inter-rater reliability of 0.65, compared to the Bartynski grades inter-rater reliability of 0.70. Discussion between observers regarding the LRD suggested difficulties consistently locating the medial end of the joint line, often obscured by osteophytosis, was the potential cause of the lower reliability. Also, the measurement often did not seem to relate to the risk of transiting nerve root compression, with an lateral recess measurement of 0 mm not necessarily relating to entrapment of the nerve root which would escape compression between the disc and superior articular facet by medial displacement into the canal.

The similarity of Bartynski grade performance in separating NC claudication patients from LBP controls compared to the measurements of the central canal was surprising and suggests that nerve root compression in the lateral recesses is potentially as important in NC symptom generation as general cauda equina entrapment in the central canal. This performance is also weak evidence that qualitative nerve root entrapment is in fact the cause of NC symptoms given the significant out performance of the LRD measurement, the latter correlated to general degenerative change and hence other potential sources of pain generation, but not particularly specific for the presence of transiting nerve root entrapment. Some caution is necessary in the above interpretation however, due to the close association between the odds of central and lateral recess nerve root entrapment.

Measurements of the neural exit foramina diameter and Lee grade were also smaller in NC patients compared to the LBP group, both in terms of their mean and most severe values per participant. In a reverse of the trend seen for the lateral recess, the performance of the NFD measurements for separating NC from LBP participants was superior to that of the Lee grade. The former achieved AUCs of 0.67 and 0.72 for the minimum and mean measurement respectively, while the latter had AUCs of 0.60 for both mean and maximum grades per participant. The difference in AUCs between NFD measurements and Lee grades was significant when comparing with the mean NFD but did not reach significance for the minimum NFD.

The reasons for the poor performance of the Lee grade relative to the NFD are not clear. It's reported inter- and intra-rater reliability exceeded those of the NFD and other tested grading systems for foraminal nerve root entrapment, though observers did report ongoing difficulty with regards to being certain exactly when the nerve root had lost its surrounding epidural fact. The Lee grade was measured on sagittal images, while the NFD was measured on axial images. It is perhaps possible that sagittal images cutting through the mid cross-section of the foramina do not well represent any potential compression of the nerve root, particularly given the degenerative pathology may be mostly occuring at the medial aspect of the foramina adjacent to the zygapophyseal joint — this remains speculation. Despite this, the similar performance of the NFD in separating LBP and NC participants to measurements of the central canal again highlights the potential importance of sites of nerve root compression outside the central canal in the generation of NC symptoms. Note, caution in interpretation is again required due to the significant degree of correlation between canal narrowing at different sites, which may potentially confound the observed relationships.

Combining measurements

In Chapter 8, two main approaches to combining the multiple different measurements available for each participant were explored. Given the potential role in symptom generation in NC of multi-site nerve root entrapment, combining information on all potential sites of entrapment was thought to have the potential to improve the so far unimpressive accuracy at which the investigated measurements could separate NC patients from the LBP control group.

The first approach tried training several different machine learning classifiers, selecting the best performing 2 models during training, and testing its classification accuracy on data withheld from the training data. A neural network and a random forrest based model were selected after showing good training performance, with small improvements in classification performance over the single measurements explored above during the training phase. This performance was was not however maintained during the test phase, with both showing drops in performance to levels where there was no significant difference over the best performing single measurements. This fall in performance from training to test phase likely represents overfitting, a common problem in machine learning that is often difficult to avoid, particularly on smaller data sets.

While this attempt to improve classification accuracy failed, to my knowledge there was no published paper attempting to combine measurements from the MRIs of LSS or NC patients in this way, at least as of early 2019 (though studies of automated grading of stenosis using computer vision methodologies exist — Q. Zhang et al. (e.g. 2017)).

The second approach presented in Chapter 8 involved designing a novel algorithm to model individual nerve root compression. This algorithm was implemented in Python and tested all measurements available for a given participant to detect which were suggestive of nerve root compression using stenosis definitions provided for each measurement. The al-

gorithm then counted the number of nerve roots compressed for that participant based on this information and that number was then explored for its ability to separate NC from LBP participants.

The correct set of thresholds to define stenosis for each measurement was not known, so large numbers of different combinations of thresholds for each measurement across the range of likely thresholds were tried, with a final set of definitions chosen that produced the best median classification accuracy for each definition. The chosen definitions of stenosis were not dissimilar to the optimum thresholds discussed for single measurements above, and included a DS-CSA of less than 60 mm^2, a Schizas grade of C or above, a Bartynski grade of 2 or above, a LRD of 4 mm or below and a NFD of 5 mm and below.

Different ways of combining information from both quantitative and qualitative measurements were also assessed, with the best overall performance provided by a rule that stated that both the quantitative and qualitative measurements at a given site had to indicate stenosis in order for a nerve root to be considered compressed. Interestingly this rule for combining measurements, coupled with the selected definition of stenosis for the Lee grade as being grade 0 or above, effectively excluded the Lee grades from the model, perhaps unsurprising given its poor performance when assessed alone.

The best combinations of stenosis definitions for use on the algorithm produced an AUC of 0.78, both on the training data and a withheld data test-set had not been used while choosing the stenosis definitions. This performance was higher than that of any of the single measurements discussed above, but this difference did not reach statistical significance. A threshold of 2 or more nerve roots predicted to be compressed by the algorithm with the definitions of stenosis given above produced a sensitivity of 80% and specificity of 62% for the presence of NC symptoms.

So why did combining measurements fail to improve upon the classification performance of single measurements?

Unlike standard statistical tests, there is no easy way to predict the required sample size for training a machine learning algorithm. The evidence of overfitting in the machine learning models produce suggests that the number of weakly predictive variables included in the machine learning models was likely too large for the given sample size, and this may have contributed towards a poor performance.

Our custom algorithm which aimed to count the number of compressed nerve roots relied upon a number of assumptions, including that the threshold of a given measurement to indicate nerve root compression was static over different spinal levels, something which is unlikely to be true for the quantitative measurements, as has already been discussed. It is possible that an algorithm that allowed the definitions of stenosis to be individually selected for each spinal level would have better performance.

It is also possible that the AUCs found for the single measurements are simply the highest possible for predicting NC symptoms from inspection of an MRI. If some threshold level of degenerative change can cause NC symptoms at a single level, combining measurements from different vertebral levels beyond the most severe level may be redundant. In addition, many of the measurements within a spinal level are closely correlated and therefore knowing one provides a lot of information about the potential values of the other measurements at that level. Any one measurement of degeneration at that level may be therefore able to predict the presence of symptoms, even if the actual cause of the symptoms is distant to the site of that particular measurement within that vertebral level.

A further assumption laying at the heart of the whole presented work, and not just this section, is that measurements based on supine MRI can accurately predict the risk of nerve root compression in the upright extended positions in which NC is generally experienced. This is discussed in detail in the section on general limitations.

Aside: The presented data does not support the multilevel compression hypothesis

As was discussed in § 1.6.3, R. W. Porter (1996) suggested a hypothesis for symptom generation in NC based upon multilevel low level compression causing venous congestion. The presented observational study presented an opportunity for collecting evidence with regards to this hypothesis.

In Chapter 4 the number of levels per spine meeting 4 different thresholds of central canal stenosis were counted. The thresholds used included: a DS-CSA less than 70 mm^2, a DS-CSA less than 100 mm^2, a Schizas grade greater than B, and a Schizas grade grater than C. If two or more levels of canal stenosis (and associated nerve root compression) was important in symptom generation, it would be expected that the number of stenotic levels would be more closely related to the presence of NC symptoms than measurements of compression at a single site. Despite this the number of centrally stenotic levels counted using any of the above thresholds did not produce AUCs significantly different to that for the minimum DS-CSA or maximum Schizas grade per participant. The number of levels considered centrally stenosed also had no correlation with any measurement of clinical severity.

It remained a possibility that the required multilevel compression was occurring outside the central canal, in the lateral recesses and neural exit foramina, and hence not detected by the above method. The possibility that this was the case was assessed using the novel algorithm for counting the number of compressed nerve roots per spine, discussed in the previous section. Instead of counting nerve roots predicted to be compressed in at least one site, the algorithm was instead used to count nerve roots where there was predicted compression in two or more sites. The performance of this version of the algorithm

across all threshold combinations was generally equivalent to or poorer than that of the algorithm when counting nerve roots predicted to be compressed at one or more sites. This again suggested that multilevel stenosis of the canal does not appear to be more closely related to NC symptom generation than single level stenosis.

These findings are in stark contrast to those presented by Hamanishi et al. (1994) who reported 89% of patients with NC had more than one level with a DS-CSA of less than 100 mm^2 while this was present in only 10% of controls. Hamanishi's sample sizes were much smaller than the current work — 53 NC patients and 39 control patients with symptomatic lesions outside the lumbar spine — and consisted purely of secondary care patients. These differences in patient population may go towards explaining some of the difference in results.

It is also of note that several other studies have found evidence that goes against what would be expected if Porter's hypothesis were true, including two papers that found less severe symptoms in patients with multi-level central stenosis compared to those with single level stenosis (Sigmundsson et al. 2011; Kuittinen et al. 2014c).

While I cannot say definitively that venous congestion secondary to multi-level compression does not contribute to NC symptom generation, the evidence presented within this thesis overall does not support this hypothesis.

9.1.2 Is there a relationship between spinal canal size on MRI and severity of clinical presentation in symptomatic LSS?

As discussed in § 1.7.2, a relationship between canal size and severity of symptoms in symptomatic spinal stenosis is expected given symptoms are thought to be caused by postural based nerve compression. Smaller canal sizes should lead to reduced requirements for dynamic change in canal dimensions before nerve root compression should occur and should reduce the effectiveness of any strategy the patient might adopt to relieve the symptoms, such as adoption of a stooped posture. Correlation between a given measurement and clinical severity is also a potentially helpful clinical feature: symptomatic LSS patients are often elderly with multiple comorbidities (Cummins et al. 2006) and hence could have multiple potential conditions all contributing towards disability. Similarly, it is not unknown for patients to be suspected to have both vascular and spinal related causes of claudication at the same time (Katz 2008). Hence, a measurement that could aid assessment of how much any canal narrowing the patient has is contributing to their mix of symptoms and disability would be potentially useful for treatment planning.

The scoping review presented in § 2.1 found inconsistent reports of relationships or lack thereof between the various measurements of anatomical LSS and a large number of different clinical severity scores. Based upon this, the DS-CSA was selected as the focus of a formal systematic review into the relationship between canal narrowing and patient pain, disability and quality of life. It was selected as it was the most commonly investigated measurement of anatomical stenosis found within the scoping review and the subject of multiple papers with relatively homogenous methodlogy yet which presented conflicting results with regards to it's relationship with clinical severity scores.

17 papers were included in the systematic review. Overall study quality was generally judged as poor, with the included studies often not providing evidence as to the reliability of their DS-CSA measurement, failing to given information on blinding of MRI observers to the results of the clinical severity scores and, perhaps most significantly, most used radiological criteria as part of their inclusion criteria often without stating the exact definition of stenosis used to include patients. All included papers also used surgical or other secondary care patient cohorts.

17 different metrics of clinical severity were assessed in the included papers. The best evidence was found for a relationship with maximum walking distance. A slight majority of papers found a positive relationship between minimal DS-CSA and maximum walking distance, but the overall pooled correlation co-efficient was not significant with substantial statistical heterogeneity. A subgroup model pooling a group of papers using patient self report found a small but significant positive correlation, similar in size to that found by Lau et al. (2017), who mixed self-report and community GPS recording and whose paper also received a fair rating for quality. This was in contrast to a group of papers using treadmill assessment of maximum walking distance which showed markedly heterogenous results and no overall significant pooled correlation. While in general, treadmill walking distance is known to have good test-retest reliability (Deen et al. 2000), it is possible variations in exact treadmill protocol played a role in the variability of the treadmill paper's findings. Treadmill based assessment does also not always correlate well with patient perception of their degree of limitation. The differences in results between the subgroups may in part be explained by the difference between capacity and performance, the former representing an individuals ability to perform in a controlled setting, and the latter performed by the individual in a day to day basis in the context of their own life (Ustün et al. 2003). These are known to be related, but measurably different, in individuals with lumbar spinal stenosis (Conway et al. 2011) and there is some evidence that self reported maximum walking distance has a stronger relationship with GPS recorded community walking distance than laboratory based measurements (Tew et al. 2013).

Repeated investigation of the relationship to the ODI, SF-36 and VAS pain scores found no correlation with the DS-CSA among the papers included in the review.

The methodological advantage of the study presented within this thesis has over the previously reviewed literature is the inclusion of community based patients and the avoidance of radiological inclusion criteria, preventing exclusion of patients with less severe symptoms, which may result in obscuration of any relationship between disease severity and clinical severity. The presented observational study was also substantially larger than the studies included in the systematic review, the largest of which included 150 patients (Burgstaller et al. 2016), opposed to the 354 BOOST participants who had MRI studies available.

Despite these advantages, no correlation was identified between any measurement of canal or dural sac size or qualitative nerve root entrapment within the central canal, lateral recess or neural exit foramen with any assessed clinical severity score. No correlation was also found between the predicted number of nerve roots compressed per participant (as judged by the algorithm introduced in Chapter 8) and any clinical severity score.

The lack of the relationship between measurements of canal/dural sac size and clinical severity variables is perhaps not unexpected, given the findings of the systematic review discussed above. However it does however raise questions as to why no relationship exists despite the expected greater risk of nerve compression at more significant degrees of stenosis. At least some of the reasons for this likely lays in the limitations of the various ways of measuring and defining canal stenosis on imaging, raised in the previous section. It is also possible that numerous other variables contribute to a patients disability and obscure a possible relationship to canal size, including psychosocial factors such as pain sensitivity, which has been found to be important in mediating reported symptom severity in symptomatic LSS (H.-J. Kim et al. 2013), and similarly pain catastrophising behaviours (maladaptive coping strategies with an exaggerated negative response to pain) have been associated with greater disability (H.-J. Kim et al. 2015). A source of obscuration for consideration is the several weak correlations identified between demographic factors and clinical severity scores discussed in Chapter 3. Despite this, it seems however likely that if a relationship between canal/dural sac size and symptom severity is present, its effect is dwarfed by the factors above.

Some associations between MRI findings and clinical severity were seen outside the outside the measurements of canal/dural-sac size. Paradoxically, there appeared to be lower levels of neuro-ischaemic symptoms (as judged by the ZCQ neuro-ischaemic domain) in patients with significant disc herniations compared to those without. Similar weak negative correlations were identified between the number of herniations per participant and the ZCQ neuro-ischaemic domain subscore and the ZCQ symptom subscale. These findings are of doubtful validity: inter-rater reliability for the presence of significant disk herniations was very poor, κ of 0.71, the correlations were weak, and p values for the grouped comparisons were just under the threshold for significance.

Osteoporotic wedge fractures were common in the BOOST participants, being present in 17% of participants. Participants with fractures showed higher pain and disability, with significantly higher ZCQ pain subdomain scores, reduction in walking distance both to symptom onset and maximal tolerable walking distance and an overall associated reduction in the EQ-5D. This latter finding is reminder that a patients presentation may be multifactorial particularly in an elderly comorbid age group, and vigilance for other potential causes of the patients presentation should also be borne in mind.

9.2 Limitations

There were a number of general limitations to the presented observational study beyond those directly related to the measurement of the spinal canal and associated structures presented above.

All MRI data used within the presented study has come from standard supine MRI studies. These are performed with the knees bent in order to improve patient comfort, which improves image quality by helping patients remain still for the scan. This position flexes the spine and reduces gravitational loading, both of which increase canal dimensions (B. I. Danielson et al. 1998; Kubosch et al. 2015). The DS-CSA measured on such supine MRI may therefore not accurately reflect the degree of stenosis present when the patient is in the standing position with the lumbar canal extended, which is when NC symptoms characteristically occur. This may represent the reason, above and beyond the previously discussed limitations of the various measurements, why the studied measurements of stenosis do not clearly differentiate NC patients from LBP and why no relationship with symptom severity is identified.

Against this, while the size of the canal may change under axial loading and an upright posture, it is likely to do so in a manner which reflects the degree of degenerative stenosis already present and demonstrable on supine MRI. Kanno et al. (2012) found that the DS-CSA on axial loading had a Pearson's correlation with the non-axially loaded DS-CSA of 0.88, implying a strong relationship between such measurements. Despite this, it is hard to argue that the additional information provided by standing MRI may not be of some clinical relevance. Willén et al. (2001) in a study of 172 patients with a mixture of low back conditions found axial loading demonstrated additional information potentially relevant to treat-

ment in 29% of patients. Ozawa et al. (2012) comparing symptomatic spinal stenosis and degenerative spondylolisthesis patients groups found that while there was no significant difference in the DS-CSA between groups on conventional imaging, axial loading induced significantly larger changes in the spondylolisthesis group. Splendiani et al. (2014) found that upright MRI could demonstrate foraminal stenosis that was not visible on supine imaging.

It may be that the writing will eventually be on the wall for standard supine MRI of the spine for suspected canal stenosis, but these techniques currently remain fairly new and have not reached widespread adoption or availability within the NHS. Until such time as they have, a definition of stenosis based on the best available data on the clinical relevance of supine MRI based measurements will retain its value.

Further limitations to the interpretation of the presented observation study are the result of the make up of the NC and LBP participant groups.

The recruited control group were referred for MRI studies by their general practitioners based on the presence of low back pain. In order to ensure no patient with potential NC was included, all patients with pain referring into the buttocks or legs were excluded in keeping with the definition of NC provided by the NASS guidelines. The decision to use a low back pain cohort was based largely on pragmatic grounds: these patients were going to be scanned anyway and hence no research dedicated scan time was required and patient recruitment was made easier.

It could be argued that the control group should have been asymptomatic volunteers. Patients with non-specific low back pain may not have completely normal lumbar spines and hence may not be an appropriate comparison group. Against this, back pain is very common, with 1 month prevalence being up to 20% (Hoy et al. 2012) in the 40 to 80 age bracket, meaning truly asymptomatic individuals are rare and themselves may not be reflective of the general population. In addition, low back pain is very common as a superimposed symptom in individuals with symptomatic lumbar spinal stenosis (de Schepper et al. 2013) and hence comparing to an asymptomatic population potentially risks finding MRI features predictive for low back pain that are not specific to NC.

The NC population was recruited as part of the BOOST RCT which is investigating the effect of a physiotherapy intervention on such patients. The trial had a pragmatic design and NC was defined for the purposes of the RCT on the basis of a series of questions, previously described in § 3.2.5. These questions were designed around the systematic review published by de Schepper et al. (2013) which found the features assessed had good sensitivity and specificity for a final diagnosis of symptomatic LSS. Beyond this, the inclusion and exclusion criteria did not make any further stipulation about the exclusion of other potential causes of claudication type pain for the participants. Vascular claudication can however be clinically difficult to destinguish from NC, may sometimes coexist alongside neurogenic claudication symptoms, and when there is clinical doubt further tests such as the ankle-brachial pressure index can be of help in ruling it out (Katz 2008). The failure to perform such tests as part of the BOOST eligability criteria means that there is a risk the BOOST participant group included some individuals with vasular claudication or a mixed picture. This may in part contribute to the poor performance of the various measurements of the canal and dural-sac in separating our case and control groups and correlating with symptom severity.

A final concern with regards to the included participant groups comes from the separate recruitment for both the NC and LBP groups. The BOOST participants were recruited at multiple centers across the UK, whereas the LBP group were recruited at a single center. Significant but small differences in mean height and ethnicity between groups was identified consistent with the source populations for the two patient groups not being completely identical. The effect of these demographics differences likely to be small given no significant difference in canal size between ethnicities was identified within the study and the mean height difference between groups was only 3 cm, however the presence of other unexplored differences between the groups cannot be excluded. Furthermore, the separate recruitment of the NC and LBP groups also limits interpretation due to potentially differing study protocols employed by such sites. Changing MR study parameters or patient positioning, such as the size of the foam wedge placed under the patients knees, could potentially effect the depiction of the canal and introduce or mask differences between the study population. The risk of this limitation is mitigated by the relatively uniform scanning protocol adopted for the lumbar spine across NHS sites, and effort made to ensure relatively consistent scanning parameters for patients undergoing studies specifically for the purposes of the trial, but this remains a limitation of the presented data.

9.3 Final conclusions and recommendations

What is the best way to measure narrowing of the spinal canal on MRI?

In § 9.1.1, the available evidence with regards to how well the investigated measurements and grading schemes separate NC from LBP participants was reviewed. No measurement stood out as being particularly strong in this regard, with all having substantial overlaps in their distributions in NC and LBP participants and resultant unimpressive AUC values. Several measurements however under-performed re-

lative to the others, including the APD, the LRD and the Lee grading system, and hence these can be excluded from any recommendation for clinical use.

In addition, the various investigated methods for combining measurements did not provide significant improvements in classification performance. Given that these methods added significant complexity without having any clear benefit, at least in terms of association of the results with symptom status of the patient, such approaches can also be excluded from any recommendations, at least in their current form.

Practicing clinicians expressed a preference for qualitative grading systems over quantitative measurements in the Delphi survey published by Mamisch et al. (2012). Such a preference makes sense from a pragmatic radiology reporting perspective as such systems can be learnt and then applied quickly, as opposed to the quantitative measurements which are slow to measure due to the need to use digital measuring tools for their assessment. Qualitative grading schemes also do not suffer problems regarding the need to adjusting for the changing amount of neural tissue in the canal and are thought to be robust to slice angulation. Given this, it makes sense to select the Schizas grade over the DS-CSA, leaving three remaining measurements, the Schizas grade of central canal stenosis, the Bartynski grade of lateral recess nerve root entrapment, and the neural exit foramen diameter.

The two qualitative grading schemes were found to have acceptable inter and intra-rater reliability during the current study (table 3.5) but the NFD had an inter-rater reliability of 0.50 and intra-rater reliability of 0.67, well below the ideal, however no alternative acceptable measurement of formainal stenosis was available to replace the NFD.

Given the possible inclusion of vascular claudication patients within the NC participant group it makes sense to prioritise specificity over sensitivity when choosing thresholds to act as a definition of stenosis. This also seems more likely to be helpful clinically: the clinician asking for an MRI in suspected symptomatic LSS is faced with a patient who has likely neurogenic claudication symptoms, and therefore wants to know if they are both attributable to the lumbar spine and caused by a vertebral level which could potentially be the target of intervention, surgical or otherwise.

Given this a Schizas grade of C or above was chosen as the recommended definition of anatomical central canal stenosis given its specificity of 85.7% for the NC participant group when using the maximum measurement (sensitivity 48.4%).

Similarly, a Bartynski grade of 3 (no CSF or other space visualised surrounding the transiting nerve root within the lateral recess) is recommended as the definition of anatomical lateral recess stenosis. This grade was found to have a specificity of 84.1% and sensitivity of 57.4% for presence of NC symptoms using the maximum measurement.

Finally a NFD measurement of 4 mm or under is proposed as a definition of anatomical foraminal stenosis and this measurement which achieved a specificity of approximately 88.9% and sensitivity of 36.8% for the presence of NC symptoms group using the minimum measurement.

While the presented study improves in some aspects of the pre-existing literature, it has its own significant limitations as discussed above, and much additional work will be required to truly clinically validate the use of the recommended measurements and associated definitions of stenosis. It is hoped that despite these limitations, the presented work does provides a starting evidence base for the use of these measurements that is at least free of the previously discussed biases present in much of the literature relating to imaging in lumbar spinal stenosis created by the lack of an accepted and truly clinically relevant reference standard. It however seems clear that development of more rigorous techniques to deal with this lack of a reference standard for imaging investigations in symptomatic LSS will be essential if a consensus is to form around an definition of anatomical LSS.

9.4 Further work

The following future work is planned by the author related to the work presented in this thesis:

- A subgroup analysis of the BOOST physiotherapy treatment effect is planned based upon several prespecified MRI based subgroups. This was originally planned to be included in the thesis but was unfortunately delayed due to the 2020 COVID-19 pandemic.
- A longitudinal study is planned, aiming to asses the relationship with the various MRI measurements and measurements of symptom severity over the course of the BOOST follow up period. This will include an assessment of whether MRI can predict those patients who subsequently went on to have surgical management of their stenosis.
- The initial plan for the thesis included assessment of a computer vision system for grading spinal stenosis developed in collaboration with Warwick Universities computer science department. The tool for training this computer system has been created but there was unfortunately not time to include this work within the thesis.
- Further development of the novel algorithm described in Chapter 8 is planned, with the intention of finding if allowing thresholds for judging nerve root compression to vary by vertebral level improves the accuracy by which it can separate NC from LBP participants.

Appendices

Appendix A

Measurements of the spinal canal

A.1 Normal values for quantitative measurements

Table A.1: Table of evidence relating to normal values of the AP diameter on CT imaging. SD – standard deviation

Method/Level	Population	Value (mm)	Citation
Axial Pedicular	N = 30 Holland No age data	L4 Mean 14.4 L4 SD 1.4 L5 Mean 15.7 L5 SD 2.8	Wilmink 1988
Axial Pedicular	N = 40 France Age: 18–40	L4 Mean 17 L4 SD 2.3 L4 Range 13–24 L5 Mean 18 L5 SD 2.4 L5 Range 13–23	Gouzien 1990
Axial Pedicular	N = 39 Spain Mean age: 38	L3 Mean 15.8 L3 SD 2.8 L4 Mean 16.3 L4 SD 3.7 L5 Mean 17.4 L5 SD 4.0 S1 Mean 18.0 S1 SD 4.6 mm	Santiago 2001
Axial Pedicular	N = 300 Egypt Age: 18–78	L1 Mean 16.8 L1 Range 12.6–26.1 L2 Mean 15.9 L2 Range 12.1–20.4 L3 Mean 15.1 L3 Range 11.9–20.4 L4 Mean 15.5 L4 Range 11.8–22.7 L5 Mean 16.4 L5 Range 11.0–24.2	Aly 2013

Table A.2: Table of evidence relating to the normal values of the AP diameter on MRI imaging. SD – standard deviation

Method/Level	Population	Value (mm)	Citation
Sagittal Mid-Vertebra	Sub-sample of 75 children Age: 15–17 W. Europe	L1 Mean 17.6 L1 SD 1.8 L2 Mean 16.4 L2 SD 1.8 L3 Mean 15.5 L3 SD 1.5 L4 Mean 16.4 L4 SD 1.6 L5 Mean 17.4 L5 SD 2.3 S1 Mean 15.0 S1 SD 2.5	Knirsch 2005
Sagittal Minimum	N = 100 China Age: 20–69	L1 Mean 15.5 L1 SD 1.5 L2 Mean 14.5 L2 SD 1.4 L3 Mean 13.7 L3 SD 1.6 L4 Mean 13.8 L4 SD 2.0 L5 Mean 14.1 L5 SD 2.1 S1 Mean 12.5 S1 SD 3.2	Cheung 2014
Axial Pedicular	N = 15 USA Age: 41–55	L2 Mean 16.3 L2 SD 2.1 L3 Mean 17.5 L3 SD 1.7 L4 Mean 18.0 L4 SD 2.7 L5 Mean 18.8 L5 SD 3.6	Singh 2005
Axial Pedicular	N = 100 Chinese Age: 20–69	L1 Mean 21.8 L1 SD 2.5 L2 Mean 21.9 L2 SD 2.5 L3 Mean 22.4 L3 SD 3.0 L4 Mean 20.2 L4 SD 2.9 L5 Mean 19.6 L5 SD 2.9 S1 Mean 12.9 S1 SD 7.8	Cheung 2014
Axial Disc	N = 20 Korean Age: 22–30	L3/4 Mean 25 L3/4 SE 0.7 L4/5 Mean 22 L4/5 SE 0.6	Chung 2000

Table A.3: Table of evidence relating to normal values of the inter-pedicular distance on CT imaging. SD – standard deviation

Method/Level	Population	Value (mm)	Citation
Axial Pedicular	N = 30 Holland No age data	L4 Mean 22.2 L4 SD 1.7 L5 Mean 25.4 L5 SD 3.6	Wilmink 1988
Axial Pedicular	N = 40 France Age: 18–40	L4 Mean 26 L4 SD 3.7 L4 Range 20–37 L5 Mean 31 L5 SD 4.6 L5 Range 21–42	Gouzien 1990
Axial Pedicular	N = 300 Egypt Age: 18–78	L1 Mean 23.8 L1 Range 17.0–30.6 L2 Mean 24.3 L2 Range 17.1–34.3 L3 Mean 25.7 L3 Range 19.1–36.7 L4 Mean 27.3 L4 Range 18.0–39.8 L5 Mean 31.5 L5 Range 21.1–43.4	Aly 2013
Axial Pedicular	N = 39 Spain Mean age: 37.9	L3 Mean 23.5 L3 SD 3.0 L4 Mean 23.7 L4 SD 3.5 L5 Mean 27.1 L5 SD 3.8 S1 Mean 31.2 S1 SD 4.0	Santiago 2001

Table A.4: Table of evidence relating to normal values of the inter-pedicular distance on MRI imaging. SD – standard deviation

Method/Level	Population	Value (mm)	Citation
Axial Pedicular	N = 100 China Age: 20–69	L1 Mean 22.5 L1 SD 1.9 L2 Mean 22.8 L2 SD 1.9 L3 Mean 24.0 L3 SD 1.9 L4 Mean 25.3 L4 SD 2.2 L5 Mean 30.0 L5 SD 2.8 S1 Mean 25.2 S1 SD 14.8 S1 SD 5.5	Cheung 2014
Axial Pedicular	N = 15 USA Age: 41–55	L2 Mean 21.9 L2 SD 1.4 L3 Mean 24.3 L3 SD 2.8 L4 Mean 23.7 L4 SD 2.4 L5 Mean 26.0 L5 SD 4.2	Singh 2005

Table A.5: Table of evidence relating to normal values of the inter-facet diameter. SD – standard deviation

Method/Level	Population	Value (mm)	Citation
CT Axial Disc	N = 40 France Age: 18–40	L3/4 Mean 22 L3/4 SD 2.9 L3/4 Range 16–28 L4/5 Mean 24 L4/5 SD 3.7 L4/5 Range 17–31	Gouzien 1990
CT Axial Disc	N = 39 Spain Mean age: 37.9	L3/4 Mean 19.9 L3/4 SD 3.8 L4/5 Mean 22.3 L4/5 SD 4.7 L5/S1 Mean 27.0 L5/S1 SD 5.3	Santiago 2001

Table A.6: Table of evidence relating to normal values of the vertebral foramen cross-sectional area. SD – standard deviation; SE – standard error

Method/Level	Population	Value (mm^2)	Citation
CT Axial Pedicular	N = 40 M France Age: 18–40	L4 Mean 288 L4 SD 60 L4 Range 188–434 L5 Mean 350 L5 SD 69 L5 Range 215–524	Gouzien 1990
MRI Axial Disc	N = 20 Korean Age: 22–30	L3/4 Mean 368 L3/4 SE 18 L4/5 Mean 363 L4/5 SE 29	Chung 2000
MRI Axial Pedicular	N = 15 USA Age: 41–55	L2 Mean 258.8 L2 SD 36.3 L3 Mean 275.3 L3 SD 24.4 L4 Mean 282.7 L4 SD 32.9 L5 Mean 323.0 L5 SD 68.5	Singh 2005

Table A.7: Table of evidence relating to normal values of the inter-ligamentous diameter. SD – standard deviation

Level	Population	Value (mm)	Citation
CT Axial Disc	N = 30 Holland No age data	L3/4 Mean 10.2 L3/4 SD 2.0 L4/5 Mean 14.6 L4/5 SD 3.8 L5/S1 Mean 21.5 L5/S1 SD 4.5	Wilmink 1988
CT Axial Disc	N = 40 France Age: 18–40	L3/4 Mean 12 L3/4 SD 2.0 L3/4 Range 9—20 L4/5 Mean 15 L4/5 SD 3.7 L4/5 Range 10—25	Gouzien 1990

Table A.8: Table of evidence relating to normal values of the dural-sac AP diameter. SD – Standard Deviation; SE – Standard Error

Level	Population	Value (mm)	Citation
MRI Axial Pedicular	N = 100 China Age: 20–69	L1 Mean 16.0 L1 SD 1.8 L2 Mean 15.4 L2 SD 1.9 L3 Mean 15.0 L3 SD 2.1 L4 Mean 13.6 L4 SD 2.1 L5 Mean 13.4 L5 SD 2.6 S1 Mean 8.3 S1 SD 5.5	Cheung 2014
MRI Axial Disc	N = 20 Korea Age: 22–30	L3/4 Mean 15 L3/4 SE 0.5 L4/5 Mean 14 L4/5 SE 0.7	Chung 2000

Table A.9: Table of evidence relating to normal values of the dural-sac transverse diameter. SD – Standard Deviation; SE – Standard Error

Level	Population	Value (mm)	Citation
CT Axial Disc	N = 30 Holland No age data	L3/4 Mean 19.0 L3/4 SD 2.8 L4/5 Mean 19.0 L4/5 SD 1.9 L5/S1 Mean 17.4 L5/S1 SD 3.8	Wilmink 1988

Table A.10: Table of evidence relating to normal values of the dural-sac cross-sectional area. SD – Standard Deviation; SE – Standard Error

Method	Population	Value (mm^2)	Citation
CT Axial Disc	N = 40 France Age: 18–40	L3/4 Mean 155 L3/4 SD 36 L3/4 Range 87–217 L4/5 Mean 157 L4/5 SD 41 L4/5 Range 90–230	Gouzien 1990
CT Axial Infra-Pedicular	N = 40 France Age: 18–40	L3/4 Mean 169 L3/4 SD 35 L3/4 Range 103–244 L4/5 Mean 168 L4/5 SD 43 L4/5 Range 94–269	Gouzien 1990
MRI Axial Disc	N = 20 Korea Age: 22–30	L3/4 Mean 260 L3/4 SE 16 L4/5 Mean 262 L4/5 SE 24	Chung 2000

Table A.11: Table of evidence relating to normal values of the lateral recess depth. SD – Standard Deviation; SE – Standard Error

Method	Population	Value (mm)	Citation
CT Lateral recess depth Infra-pedicular	N = 300 Egypt Age: 18–78	Mean 6.7 Range 4–14	Aly 2013
CT Lateral recess depth Infra-pedicular	N = 39 Spain Mean age: 37.9	L4 Mean 6.3 L4 SD 0.22 L5 Mean 5.5 L5 SD 0.16 S1 Mean 5.6 S1 SD 0.14	Santiago 2001
MRI SSDC Disc	N = 20 Korean Age: 22–30	L3/4 Mean Right 4.3 L3/4 SE Right 0.28 L3/4 Mean Left 4.4 L3/4 SE Left 0.22 L4/5 Mean Right 3.4 L4/5 SE Right 0.30 L4/5 Mean Left 3.2 L4/5 SE Left 0.25	Chung 2000

Table A.12: Table of evidence relating to normal values of the neural exit foramen diameter. SD – Standard Deviation; SE – Standard Error

Level	Population	Value (mm)	Citation
CT Axial Infra-pedicular	N = 39 Spain Mean age: 37.9	L4 Mean 10.3 L4 SD 2.5 L5 Mean 9.9 L5 SD 1.9 S1 Mean 8.8 S1 SD 3.1	Santiago 2001

Appendix B

Literature Review

B.1 Scoping review

B.1.1 Scoping review methodology

Scoping Review Aims

The scoping review aims to assess the range of published literature on the relationship between findings on MR examination of the lumbar spine and patient presentation and prognosis in LSS. Ultimately the review is aimed to guide the design of future more focused systematic reviews and experimental work.

Objectives:

- To assess the variation in study populations, both symptomatic and control, and associated inclusion and exclusion criteria.
- To achieve an overview of the degree of evidence supporting the use of the quantitative and qualitative measures of spinal stenosis described in the literature.
- To gauge whether the literature is of sufficient homogeneity to potentially allow meta-analysis of published results in a subsequent systematic review.

Search strategy

A pragmatic search strategy was chosen, balancing completeness of the search results with the ability of a single individual to complete the search in a reasonable time frame. The search was designed to return papers referring to both lumbar stenosis or neurogenic claudication and MRI examination of the lumbar spine and was performed on the MEDLINE database on the 9th November 2016. The references generated were screened by title and then abstract, with papers at each stage not meeting the inclusion and exclusion criteria detailed below being excluded. The full text of the remaining citations were reviewed by a single individual for inclusion. No review of the included papers bibliography was performed.

Inclusion Criteria:

- Case populations much consist of adults with a primary diagnosis of symptomatic degenerative lumbar spinal stenosis on basis of expert opinion.
- Papers must either:
  - Report a quantitative or qualitative measure of central spinal stenosis on MRI examination and relate this to a validated quantitative measure of pain, disability or quality of life
  - Assess the difference in a quantitative or qualitative measure of central spinal stenosis on MRI examination between a case population (as described above) and a control population.

Exclusion Criteria:

- Papers including patients with prior spinal surgery or assessing post operative imaging.
- Papers including patients with a diagnosis of non-degenerative spinal pathology, acute spinal conditions or any congenital syndrome where these patients are not analysed separately.
- Papers including patients with lateral and/or foramina stenosis without central stenosis where these patients are not analysed separately.
- Papers whose full text was not available in English either through open publication or via the journal subscriptions of the University of Warwick.

Data Extraction

Data extraction was performed by a single individual. Data was extracted primarily to a master table with columns for recording descriptions of the case and control populations, exclusion criteria, MR and outcome variables used, methodological details, study type and a summary of the papers results.

B.1.2 Scoping review results: additional tables

Table B.1: Scoping review search strategy

Number	Search Term	Citations
1.	lumbar	94 332
2.	lumbosacral	19 619
3.	sacral	12 621
4.	exp Lumbar Vertebrae/	44 696
5.	exp Lumbosacral Region/	10 861
6.	exp Sacral Region/	3523
7.	stenosis	158 912
8.	claudication	11 450
9.	exp Spinal Stenosis/	5018
10.	exp Intermittent Claudication/	7409
11.	exp Nerve Root Compression/	4317
12.	”Magnetic resonance”	575 262
13.	MRI	160 362
14.	MR	115 279
15.	exp Magnetic Resonance Imaging/	365 284
16.	1 OR 2 OR 3 OR 4 OR 5 OR 6	114 736
17.	7 OR 8 OR 9 OR 10 OR 11	172 624
18.	12 OR 13 OR 14 OR 15	643 233
19.	16 AND 17	7121
20.	18 AND 19	1732

Table B.2: Results of studies assessing the relationship between quantitative measures of the dural-sac and clinical symptoms.

Measurement	Clinical Scoring System	Reference	N	Result
Min. DS-CSA	VAS/NRS Pain	Barz 2008	28	No linear correlation.
		Sigmundsson 2011	109	No linear correlation.
		Kanno 2012	88	No linear correlation.
		Kuittinen 2014	84	No linear correlation.
		Kim 2015	117	No linear correlation.
		Burgstaller 2016	150	No linear correlation.
	ODI	Sirvanci 2008	63	No linear correlation.
		Sigmundsson 2011	109	No linear correlation.
		Kuittinen 2014	84	No linear correlation.
		Hong 2015	73	No linear correlation.
		Kim 2015	117	No linear correlation.
	Objective WD	Barz 2008	28	+ve correlation ($p = 0.003$)
		Moon 2005	35	No linear correlation.
		Kanno 2012	88	+ve correlation ($p < 0.05$)
		Kuittinen 2014	84	No linear correlation.
	Subjective walking distance	Sigmundsson 2011	109	No linear correlation.
		Kim 2015	117	+ve correlation ($p = 0.045$).
	Duration of symptoms	Kanno 2012	88	-ve correlation ($p < 0.05$)
		Kim 2015	117	No linear correlation.
	BDI	Hong 2015	73	No linear correlation.
		Kuittinen 2014	84	No linear correlation.
	SF-36	Hong 2015	73	No linear correlation.
		Sigmundsson 2011	109	No linear correlation.
	ZCQ PD	Burgstaller 2016	150	No linear correlation.
	VAS Numbness	Kanno 2012	88	No linear correlation.
	JOA Score	Kanno 2012	88	+ve Correlation ($p < 0.05$).
	EQ-5D	Sigmundsson 2011	109	No linear correlation.
	BAI	Hong 2015	73	No linear correlation.
	PSQI	Hong 2015	73	No linear correlation.
DS-CSA by Level	SSM Pain Domain	Burgstaller 2016	150	No linear correlation.
	NRS Pain	Burgstaller 2016	150	No linear correlation.
	Objective WD	Zeifant 2008	63	+ve correlation L1/2 ($p = 0.032$).
Mean DS-CSA	Objective WD	Zeifant 2008	63	No linear correlation.
Min. DS-APD	SSM Pain Domain	Burgstaller 2016	150	No linear correlation.
	NRS Pain	Burgstaller 2016	150	No linear correlation.
DS-APD by Level	SSM Pain Domain	Burgstaller 2016	150	No linear correlation.
	NRS Pain	Burgstaller 2016	150	No linear correlation.

Table B.3: Results of studies comparing quantitative measurements of the dural-sac between symptomatic LSS patients and controls.

Measurement	Reference	Case N	Control N	Results
DS-CSA	Hamanashi 1994	53	15	Mean cross sectional area of the dural tube was smaller at all vertebral levels in neurogenic claudication patients compared to age matched controls ($p < 0.01$).
	Schizas 2010	68	27	Min. DS-CSA per patient at the pedicular level was smaller in LSS patients managed surgically compared to those managed conservatively ($p < 0.0001$) and controls ($p < 0.0001$). The ratio of disc level DS-CSA to pedicular level DS-CSA was smaller in symptomatic stenosis at all levels compared to control groups ($p < 0.005$).
	Tomkins-Lane 2012	50	76	Min. DS-CSA per patient did not significantly correlate with 7 day community walking distance or 15 minute treadmill velocity.
	Kim 2015	117	91	Mean and min DS-CSA per patient was smaller in the symptomatic stenosis group compared to the control group.
DS-APD	Tomkins-Lane 2012	50	76	Min. DS-CSA per patient did not significantly correlate with 7 day community walking distance or 15 minute treadmill velocity.

Table B.4: The results of studies assessing the relationship between quantitative measures of the dural-sac and patient prognosis in symptomatic LSS.

Measurement	Reference	N	Result
Min. DS-CSA	Schizas 2010	68	Min. DS-CSA <75 mm2 or DS-CSA <160 mm2 at the pedicular level was predictive of an increased risk of failure of conservative treatment. No significant correlation was found between min. DS-CSA and ODI improvement post surgery.
	Park 2011	66	No significant difference in min. DS-CSA was found between groups improving and not improving with surgery. No significant correlation with patient improvement with surgery and DS-CSA
	Sigmundsson 2012	109	Significant negative correlation between DS-CSA and change in VAS back pain with surgery ($p < 0.05$). No correlation was found with absolute VAS back pain at one year. No significant correlation with change in VAS leg pain. A positive correlation was found with absolute DS-CSA and VAS leg pain at one year ($p = 0.03$).
	Kuittinen 2014	84	No significant correlation was found with post-operative improvement in NRS leg, back and overall pain, ODI, maximum treadmill walking distance or patient satisfaction.
Min. DS-APD	Verhoof 2008	12	No significant differences in AP diameter between those having residual or recurrent symptoms post interspinous implant insertion and those showing improvement in reported pain.

Table B.5: Results of studies assessing qualitative grading systems for dural-sac narrowing. N numbers in brackets represent control group size

Grading System	Reference	N	Results
0 — No contact; 1 — thecal sac contact, no deviation of nerve roots; 2 — thecal sac contact, deviation of nerve roots; 3 — compression of nerve roots	Alicioglu 2012	70	No correlation with JOA score or post surgical JOA recovery rate. No significant difference in grade between those improving with surgery or not.
0 — normal; 1 — anterior CSF space reduced but all cauda equina clearly separated; 2 — partial cauda equina aggregation; 3 — marked no cauda equina roots visualised separately.	Park 2014	47	No correlation with Roland score post epidural steroid injection.
0 — normal; 1 — narrowing, CSF clearly visible; 2 — little or no CSF between nerve roots.	Kuittinen 2014	84	Severe stenosis predicted less postoperative NRS pain and better post-operative satisfaction compared to moderate stenosis (p = 0.028, p = 0.029 respectively). No correlation with post-operative ODI or subjective walking distance.
	Kuittinen 2014	84	NRS pain was lower (p=0.008) and treadmill walking distance was significantly higher in severe stenosis compared to moderate. There was no significant difference in ODI, VAS overall or BDI between grades 1 and 2.
A — clear CSF visible with inhomogeneous distribution; B — nerve roots occupy the whole of the sac, some CSF still present; C — no CSF visible; D — no epidural fat visible posteriorly.	Schizas 2010	68 (27)	More severe grades were seen in the symptomatic stenosis patients compared to the control group (p <0.05). No correlation between the most severe grade of stenosis and ODI.
	Kim 2013	94	No correlation with VAS leg pain, back pain or ODI.
	Moojen 2018	159	No correlation with MRDQ, VAS leg or back pain.
	Burgstaller 2016 Weber 2016	150 202	No correlation with SSM pain domain or NRS pain. No correlation with ODI or NRS pain.

Table B.6: Results of papers investigating the sedimentation sign. Control group N values are shown in brackets.

Reference	N	Results
Barz 2010	100 (100)	Sedimentation sign was positive in 94% of patients with degenerative lumbar spinal stenosis but 0% of control patients. There were 6 false negatives. Note should be taken that the definition of case and control populations included a DS-CSA cut off.
Macedo 2013	50 (22)	Sedimentation sign was positive in 54% of people diagnosed with symptomatic stenosis, 23% of people with isolated lateral stenosis and 2% of people with posterolateral disc herniation. The difference between groups was significant (p<0.01).
Barz 2014	118	Conservatively managed patients negative for the sedimentation sign showed higher improvement with time in ODI and VAS back pain compared to those positive (p <0.01). No significant difference between in improvement of ODI and VAS back pain between positive and negative patients was seen in those managed surgically. VAS leg pain did not differ significantly between positive and negative patients for either management group.
Moses 2015	115	No significant difference in ODI, SF-36, stenosis, back pain or leg pain bothersomeness index between those sedimentation sign positive or negative. Those with a positive sedimentation sign demonstrated a larger treatment effect for surgery as judged by ODI and stenosis bothersomeness index (p = 0.023 and 0.094 respectively). There was no significant difference in surgical treatment effect for SF-36 scores between sedimentation sign positive and negative groups.

Table B.7: Results of papers investigating the redundant nerve root sign.

Reference	N	Results
Min 2008	68	No significant difference between pre-operative JOA score, post-operative JOA score, JOA recovery rate or success rate of surgery between redundant nerve root positive and negative groups. In patients who had redundant nerve roots there was significant correlation between post-operative JOA score and JOA recovery rate with the length of the redundant nerve roots. No significant correlation was seen with pre-operative JOA score.
Moojen 2018	159	No significant difference in MRQD, VAS leg or back pain in those with or without redundant nerve roots at baseline before surgery. Presence of redundant nerve predicted a higher rate of good response to surgical decompression compared to those with absent redundant nerve roots (p <0.01).

Table B.8: The results of studies comparing quantitative measurements of the spinal canal to clinical findings.

Measurement	Clinical Scoring System	Reference	N	Result
Min. VF-CSA	ODI	Eun 2012	163	No significant linear correlation.
		Kim 2015	117	No significant linear correlation.
	VAS Pain	Eun 2012	163	No significant linear correlation.
		Kim 2015	117	No significant linear correlation.
	Subjective WD	Kim 2015	117	No significant linear correlation.
Min. APD	Objective WD	Moon 2005	163	No significant linear correlation.
		Haig 2006	50	No significant linear correlation.
		Geisser 2007	50	No significant linear correlation. No significant difference between particiapants stratified by AP diameters of 13 mm or 10 mm.
	PDI	Haig 2006	50	No significant linear correlation.
		Geisser 2007	50	Participants with AP diameters <13 mm had higher PDI scores than those with AP diameters >13mm ($p < 0.05$). No significant difference between groups using a threshold of 10 mm. No linear correlation identified.
	VAS Pain	Haig 2006	50	No significant linear correlation.
	QBPDI	Geisser 2007	50	No linear correlation. No significant difference between participants stratified by AP diameters of 13 mm or 10 mm.
	MPQ	Geisser 2007	50	No linear correlation. No significant difference between participants stratified by AP diameters of 13 mm or 10 mm.
Min. LRH	ZCQ PD	Burgstaller 2016	150	No significant linear correlation.
	NRS Pain	Burgstaller 2016	150	No significant linear correlation.
Foraminal CSA	VAS Pain	Barz 2008	25	No significant linear correlation.

Table B.9: Results of studies assessing the clinical relevance of qualitative measurements of the spinal canal. Control group N is shown in brackets.

Grading System	Reference	N	Results
No — normal; mild — compromise of less than 1/3 area; moderate — compromise of 1/3 – 2/3; severe — compromise of more than 2/3	Ishimoto 2013	105 (833)	Symptomatic stenosis prevalence increased with severity of central stenosis grade (men: $p = 0.009$; women: $p = 0.004$). Severe central canal stenosis was related to symptomatic individuals but moderate stenosis was not.
	Burgstaller 2016	150	No correlation between central canal grade and SSM pain domain or NRS pain.
	Burgstaller 2016	150	No correlation between neural exit foramina grade and SSM pain domain or NRS pain when analysis was performed for all graded levels or for the most severe level for each participant.
0 — normal; 1 — mild central canal epidural fat overgrowth, 2 — moderate overgrowth; 3: — severe overgrowth	Burgstaller 2016	150	No correlation between central canal grade SSM pain domain or NRS pain.
Presence or absence of epidural fat around the central canal.	Moojen 2018	159	No difference in VAS pain or MRDQ between groups at baseline before surgery. Absence of epidural fat predicted a higher rate of satisfactory response to surgical decompression ($p = 0.03$).
Mild, moderate, severe not otherwise defined.	Kapural 2007	719	No significant correlation between central grading and VAS pain scores.

Table B.10: Results of studies assessing the clinical relevance of quantitative measurements of nerve root compression within the lateral recess and neural exit foramina.

Grading System	Reference	N	Results
0 — No contact of the disc with the nerve root; 1 — contact with no deviation of nerve roots; 2 — contact with deviation of nerve root; 3 — Compression of nerve root (nerve roots deformed).	Sirvanci 2008	63	No significant correlation of lateral recess grade with ODI.
	Alicioglu 2012	70	No correlation with lateral recess grade and JOA score or improvement in JOA score post surgery. No significant difference in lateral recess grade between those improving with surgery and those showing no improvement.
	Burgstaller 2016	50	No correlation of neural exit foramen grade with SSM pain domain or NRS pain when analysis was performed for all graded levels or for the most severe level for each participant.
0 — no narrowing; 1— narrowing, no root compression; 2 — significant narrowing, nerve root flattened, preservation of cerebrospinal fluid; 3 — severe root compression, no CSF.	Burgstaller 2016	50	No correlation of lateral recess grade with SSM pain domain or NRS pain when analysis was performed for all graded levels or for the most severe level for each participant.
0 — normal; 1 — narrowing without root compression; 2 — nerve root compression	Kuittinen 2014	84	No significant correlation between lateral recess grade and post-operative improvement in NRS leg pain, ODI, maximum objective walking distance or patient satisfaction.
0 — Normal foramina; 1 — deformity of the epidural fat but still completely surrounding the nerve root; 2 — epidural fat only partially surrounds the nerve root; 3 — obliteration of the epidural fat.	Alicioglu 2012	70	No correlation between foraminal grade and JOA score or improvement in JOA score post surgery. No significant difference in lateral recess grade between those improving with surgery and those showing no improvement.

Table B.11: The results of cross-sectional studies assessing the clinical relevance of multi-level stenosis.

Measurement	Clinical Scoring System	Reference	N	Result
Num. Stenotic Levels	ODI	Sirvanci 2008	63	No linear correlation
		Park 2010	1226	No linear correlation
		Sigmundsson 2011	109	No linear correlation
		Hong 2015	73	Positive correlation
	SF-36	Park 2010	1226	No linear correlation
		Sigmundsson 2011	109	No linear correlation
		Hong 2015	73	No linear correlation
	VAS Pain	Kapural 2007	719	No linear correlation
		Sigmundsson 2011	109	Negative correlation with leg pain (p = 0.03)
	Objective WD	Zeifang 2008	63	No linear correlation
	Subjective WD	Sigmundsson 2011	109	No linear correlation
	SBI	Park 2010	1226	No linear correlation
	LPBI	Park 2010	1226	No linear correlation
	LBPBI	Park 2010	1226	No linear correlation
	EQ-5D	Sigmundsson 2011	109	No linear correlation
	BDI	Hong 2015	73	No linear correlation
	BAI	Hong 2015	73	No linear correlation
	PSQI	Hong 2015	73	No linear correlation
Multi-Level Stenosis	Objective WD	Zeifang 2008	63	No significant relationship.
		Kuittinen 2014	84	Patients with multi-level stenosis had longer maximum WD (p = 0.022)
	SF-36	Sigmundsson 2011	109	Patients with multi-level stenosis showed higher levels of SF-36 general health domain scores (p = 0.04).
		Hong 2015	73	No significant relationship.
	ODI	Sigmundsson 2011	109	No significant relationship.
		Hong 2015	73	Higher ODI is seen in multi-level stenosis (p = 0.04)
	VAS Pain	Sigmundsson 2011	109	No significant relationship.
	EQ-5D	Sigmundsson 2011	109	No significant relationship.
	BDI	Hong 2015	73	No significant relationship.
	BAI	Hong 2015	73	No significant relationship.
	PSQI	Hong 2015	73	No significant relationship.

Table B.12: The results of studies assessing the prognostic significance of multi-level stenosis.

Reference	N	Result
Park 2010	1226	This cohort study measured SF-36, ODI stenosis, leg pain or low back pain bothersomeness index as outcome variables. Patients without degenerative spondylolisthesis (n = 619) with single level stenosis who underwent surgery showed increased overall patient satisfaction at 2 years and higher surgical treatment effects in terms of satisfaction compared to multi-level stenosis. Patients with spondylolisthesis (n = 607) showed higher overall SF-36 and ODI scores in single level patients at two years and higher surgical treatment effect in terms of low back pain bothersomeness compared to multi-level stenosis. There was no significant difference in the other outcome variables between groups.
Sigmundsson 2012	109	Patients with multi-level stenosis had lower post-operative VAS leg pain scores compared to single level disease (p = 0.06).
Kuittinen 2014	84	One level central stenosis (any narrowing of the central canal) predicted lower NRS post-operative back pain compared to patients with two or more levels (p = 0.026). No significant difference found for post-operative NRS leg or overall pain, ODI, maximum treadmill walking distance between groups.

B.2 Systematic review search strategy

The search was performed on the 23 August 2017.

Table B.13: EMBASE Classic + EMBASE 1980 to 2017 Week 34

No.	Search Term	Citations
1	lumbar	154 902
2	lumbosacral	15 795
3	sacral	23 146
4	exp Lumbar Vertebrae/	18 658
5	exp Lumbosacral Region/	488
6	exp Sacrum/	9247
7	stenos?s	270 850
8	claudication	18 911
9	Nerve root	14 196
10	Cauda equina	7284
11	Compres*	202 604
12	exp Spinal Stenosis/	9890
13	exp Intermittent Claudication/	9783
14	exp Nerve root compression/	2683
15	Magnetic resonance	1 100 330
16	MRI	336 718
17	exp Magnetic Resonance Imaging/	761 918
18	1 OR 2 OR 3 OR 4 OR 5 OR 6	184 654
19	(9 OR 10) AND 11	6373
20	7 OR 8 OR 12 OR 13 OR 14 OR 19	291 560
21	15 OR 16 OR 17	1 133 129
22	18 AND 20 AND 21	3942

Table B.14: MEDLINE (R) 1946 to August Week 2 2017

No.	Search Term	Citations
1	lumbar	99 651
2	lumbosacral	20 564
3	sacral	13 403
4	exp Lumbar Vertebrae/	47 085
5	exp Lumbosacral Region/	11 383
6	exp Sacrococcygeal region/	3666
7	exp Sacrum/	8233
8	stenos?s	174 632
9	claudication	12 316
10	Nerve root	5936
11	Cauda equina	5048
12	exp Cauda equina/	3228
13	Compres*	127 248
14	exp Spinal Stenosis/	5268
15	exp Intermittent Claudication/	7966
16	exp radiculopathy/	4493
17	Magnetic resonance	606 334
18	MRI	173 769
19	exp Magnetic Resonance Imaging/	388 420
20	1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7	123 380
21	(10 OR 11 OR 12) AND 13	3327
22	8 OR 9 OR 14 OR 15 OR 16 OR 21	191 445
23	17 OR 18 OR 19	633 567
24	20 AND 22 AND 23	2223

Table B.15: Web of Science search strategy

No.	Search Term	Citations
1	TS = (lumbar) OR TI = (lumbar)	83 887
2	TS = (lumbosacral) OR TI = (lumbosacral)	8500
3	TS = (sacral) OR TI = (sacral)	13 690
4	TS = (stenosis) OR TI = (stenosis)	129 586
5	TS = (claudication) OR TI = (claudication)	9241
6	TS = (Nerve root) OR TI = (Nerve root)	24 219
7	TS = (Cauda equina) OR TI = (Cauda equina)	3998
8	TS = (Compres*) OR TI = (Compres*)	497 070
9	TS = (Magnetic resonance) OR TI = (Magnetic resonance)	456 743
10	TS = (MRI) OR TI = (MRI)	250 070
11	#1 OR #2 OR #3	99 865
12	#6 OR #7	27 346
13	#12 AND #8	3505
14	#4 OR #5 OR #13	140 004
15	#9 OR #10	586 210
16	#11 AND #14 AND #15	1548

Table B.16: Cochrane Library search strategy

No.	Search Term	Citations
1	lumbar	9963
2	lumbosacral	1027
3	sacral	904
4	MeSH descriptor: Lumbar Vertebrae – explode all trees	2599
5	MeSH descriptor: Lumbosacral Region – explode all trees	394
6	MeSH descriptor: Sacrum – explode all trees	119
7	stenos?s	9155
8	claudication	1945
9	Nerve root	949
10	Cauda equina	130
11	Compres*	7813
12	MeSH descriptor: Spinal Stenosis – explode all trees	265
13	MeSH descriptor: Intermittent Claudication – explode all trees	850
14	MeSH descriptor: Radiculopathy – explode all trees	271
15	Magnetic resonance	19 789
16	MRI	10 060
17	MeSH descriptor: Magnetic Resonance Imaging – explode all trees	7463
18	#1 OR #2 OR #3 OR #4 OR #5 OR #6	11 118
19	(#9 OR #10) AND #11	253
20	#7 OR #8 OR #12 OR #13 OR #14 OR #19	11 200
21	#15 OR #16 OR #17	21 215
22	#18 AND #20 AND #21	158

B.3 Systematic review results: additional tables

Table B.17: Reported sensitivity (sens.) and specificity (spec.) of the sedimentation sign for the presence of LSS.

Citation	Case Population	Control Population	Sens. %	Spec. %
Barz 2010	100 orthopedic patients with NC, DS-CSA <80 mm^2 and WD <200 meters	100 orthopedic patients with non-specific LBP, DS-CSA >120 mm^2 and WD >1000 m	94	100
Tomkins-Lane 2013	67 secondary care patients with symptomatic LSS from the Michigan spinal stenosis study. Definition of radiological stenosis not provided.	31 secondary care patients with non-specific LBP, 4 with vascular claudication and 46 asymptomatic controls.	42–66	49–78
Laudato 2015	110 secondary care patients with symptomatic LSS. Definition of radiological stenosis not provided. 73 managed surgically, 37 conservatively.	27 secondary care patients with non-specific LBP.	56	93
Zhang 2017	105 orthopedic outpatients with NC and DS-CSA <120 mm^2.	215 orthopedic outpatients with non-specific LBP.	77.1	53.0

Table B.18: Reported sensitivity (sens.) and specificity (spec.) of grades C and D of the grading system proposed by Schizas et al. for the presence of LSS.

Citation	Case Population	Control Population	Sens. %	Spec. %
Schizas 2010	95 secondary care patients with symptomatic LSS. Definition of radiological stenosis not provided. 37 managed surgically, 31 conservatively.	27 secondary care patients with non-specific LBP.	61.8	88.8
Laudato 2015	110 secondary care patients with symptomatic LSS. Definition of radiological stenosis not provided. 73 managed surgically, 37 conservatively.	27 secondary care patients with non-specific LBP.	76	92

Table B.19: Case-control studies comparing measurements of the ligamentum flavum.

Citation	Case Population	Control Population	Results
Kim 2015	117 pain clinic outpatients with symptoms of NC and radiological LSS diagnosed by a neuroradiologist.	91 individuals undergoing MRI as part of a routine medical checkup without symptoms of LSS	Ligamentum flavum area and ligamentum flavum thickness at the most stenotic level were greater in the neurogenic claudication group compared with the control group ($p < 0.001$).
Kim 2017	168 pain clinic outpatients with symptoms of NC and radiological LSS diagnosed by a neuroradiologist.	167 individuals undergoing MRI as part of a routine medical checkup without symptoms of LSS	ROC analysis showed a ligamentum flavum area cut-off at L4/5 of 105.9 mm^2 had an 80.1% and 76.0% specificity with AUC of 0.83. A ligamentum flavum thickness of 3.74 mm had a 70.5% sensitivity and 66.5% specificity with AUC of 0.72.

Table B.20: Reported findings for other MRI measurements. FA – fractional anisotropy.

Citation	MRI Measurement	Case Population	Control Population	Findings
Manaka 2003	Contrast Enhanced Phlebography	58 secondary care patients with canal stenosis and claudication.	15 patients with other lumbar diseases and 13 asymptomatic volunteers.	Significantly greater venous plexus defects were seen in neurogenic claudication patients compared both to other lumbar diseases ($p < 0.0001$) and asymptomatic controls ($p < 0.0001$). More severe grades were associated with shorter claudication distances.
Hou 2015	Nerve Root Tractography	31 patients with NC and a min. DS-CSA less than two-thirds that at the corresponding normal intervertebral space.	20 healthy volunteers	Mean FA values were lower in neurogenic claudication patients ($p < 0.001$). Healthy volunteers showed symmetrical and complete tractography whereas 49% of nerve roots in neurogenic claudication participants showed thinning and distortion and 23% showed focal loss of tracing.
Kim 2015	Spinal Canal Area	117 pain clinic outpatients with symptoms of NC and radiological LSS diagnosed by a neuroradiologist.	91 individuals undergoing MRI as part of a routine medical checkup without symptoms of lumbar spinal stenosis	Minimal spinal canal area was significantly smaller in the NC group compared to the control group ($p < 0.001$).
Kobayashi 2015	Dynamic Contrast Enhancement of the Cauda Equina	23 secondary care patients with spinal stenosis (DS-CSA <70 mm^2) and NC.	10 asymptomatic volunteers with DS-CSA >100 mm^2 at all levels	In patients with stenosis, the signal intensity of the cauda equina at the site of stenosis peaked later than controls and maintained abnormal enhancement for up to 10 minutes.
Ghasemi 2016	Sagittal alignment of the lumbar spine	120 orthopedic outpatients with NC and an AP spinal canal diameter less than 9 mm.	120 orthopedic outpatients with non-specific LBP and a group of 120 patients with lumbar disc herniation and radiculopathy.	The sacral angle in NC patients was significantly smaller compared to control and lumbar disc herniation groups ($p < 0.05$). No significant difference in vertical angle of sacral curvature between groups identified.
Chun 2017	Phase Contrast Measurement of CSF Flow	4 patients with NC and anatomical LSS (not further defined).	8 healthy volunteers.	No significant difference in CSF flow between patients and controls at rest. After walking for 30 minutes or to claudication onset, claudication patients showed significantly lower velocities than controls.
Jiang 2017	Multifidus Atrophy	40 patients undergoing lumbar decompression for symptomatic LSS including NC symptoms.	40 healthy volunteers.	Multifidus fat infiltration ratio and muscle asymmetry ratio were significantly greater and the lumbar muscularity (adjusted for vertebral body size) was significantly lower in the neurogenic claudication group compared to the control group ($p < 0.05$). No significant difference in overall multifidus cross sectional area was identified between groups.

Table B.21: Table of study results relating to the Oswestry Disability Index

Reference	Study Population	Results
Sirvanci 2008	63 Surgically Managed Patients	No significant correlation between min. DS-CSA and ODI. No significant relationship between min. DS-CSA categorised as severe stenosis (< 76 mm^2) moderate stenosis (76–100 mm^2) and normal (>100 mm^2) and ODI disability categories.
Schizas 2010	68 Secondary Care Outpatients	No correlation between the min. DS-CSA per patient and ODI in either surgical (n = 37) or conservatively (n = 31) managed patients (groups were analyzed separately).
Sigmundsson 2011	58 Surgically Managed Patients	No significant correlation between min. DS-CSA or the number of stenotic levels (threshold DS-CSA < 70 mm^2) and the ODI. No significant difference in ODI between single level and multi-level stenosis groups.
Goni 2014	50 Secondary Care Outpatients	No significant relationship between min. DS-CSA categorised as severe stenosis (<76 mm^2) moderate stenosis (76–100 mm^2) and normal (> 100 mm^2) and ODI disability categories. A DS-CSA < 70 mm^2 was associated with more severe disability groups (p = 0.026).
Hong 2015	74 Referrals for Epidural Adhesiolysis	No significant correlation between min. DS-CSA and ODI. Small positive correlation between number of stenotic levels and ODI (threshold DS-CSA < 100 mm^2, p < 0.045). Patients with single level stenosis had lower ODI scores compared to those with multi-level stenosis (p = 0.040).
Kim 2015	117 Secondary Care Outpatients	No significant correlation between min. DS-CSA per patient and ODI.
Marawar 2016	30 Secondary Care Outpatients	No correlation between min. DS-CSA per patient and ODI (patients with DS-CSA < 70 and > 70 mm2 analyzed separately) or between average DS-CSA and ODI (multi-level or single level stenosis analyzed separately, threshold DS-CSA < 100 mm^2). No significant difference in ODI between min. DS-CSA < 70 and > 70 mm^2 groups or between single level,and multi-level stenosis.
Prasad 2016	48 Surgically Managed Patients	No significant correlation between min. DS-CSA per patient and ODI.
Alsaleh 2017	54 Secondary Care Outpatients	No significant correlation between min. DS-CSA per patient and ODI.
Lau 2017	70 Secondary Care Outpatients	No significant correlation between min. DS-CSA and ODI on supine or standing MRI.

Table B.22: Table of study results relating to the Short Form 36 Questionnaire.

Reference	Study Population	Results
Sigmundsson 2011	109 Surgically Managed Patients	No significant correlation between the min. DS-CSA or number of stenotic levels (threshold DS-CSA < 70 mm^2) and SF-36 physical function, role physical, bodily pain and general health domains. Patients with multilevel stenosis had significantly greater SF-36 general health scores (p = 0.04) but no difference in other domains.
Hong 2015	73 Referrals for Epidural Adhesiolysis	No significant correlation between min. DS-CSA and SF-36 bodily pain physical function role physical and general health domains. No correlation between number of stenotic levels (threshold DS-CSA < 100 mm^2) and SF-36 domains or difference between single and multi-level stenosis groups.
Marawar 2016	30 Secondary Care Outpatients	No significant correlation between DS-CSA at any analysed level and any SF-36 domain.
Prasad 2016	48 Surgically Managed Patients	No significant correlation between min. DS-CSA and SF-36 physical function and mental health domains.
Alsaleh 2017	54 Secondary Care Outpatients	No significant correlation between min. DS-CSA and SF-36 score.
Lau 2017	70 Secondary Care Outpatients	No significant correlation between min. DS-CSA and SF-36 physical health or mental health domains on supine or standing MRI.

Table B.23: Table of study results relating to pain severity.

Reference	Study Population	Results
Barz 200819	25 Surgically Managed Patients	No significant correlation between min. DS-CSA and overall pain.
Sigmundsson 201124	109 Surgically Managed Patients	No significant correlation between min. DS-CSA and leg or low back pain identified. Small negative correlation between the number of stenotic levels (threshold DS-CSA < 70 mm^2) and leg pain (r = -0.24 p = 0.03) but no correlation with low back pain was identified. No significant difference in leg or low back pain between single level and multi-level stenosis groups.
Kanno 201223	88 Referrals for Surgical Management	No significant correlation was identified between minD□S-CSA and leg pain or numbness on standard supine MRI. On axial loading a moderate negative correlation with leg numbness (r = -0.35, p < 0.01) was identified but the relationship with leg pain remained non-significant.
Hong 201528	73 Referrals for Epidural Adhesiolysis	No significant correlation between min. DS-CSA or the number of stenotic levels (threshold DS-CSA < 100 mm^2) and overall pain. No significant difference in overall pain between single and multi-level stenosis groups.
Kim 201522	117 Secondary Care Outpatients	No significant correlation between min. DS-CSA and leg or low back pain.
Burgstaller 201630	150 Secondary Care Outpatients	No significant correlation between min. DS-CSA or per segment DS-CSA and overall pain scores.
Marawar 201642	30 Secondary Care Outpatients	No significant correlation between leg pain and DS-CSA at any analysed levels. Right and left leg pain were analysed separately.
Prasad 201618	48 Surgically Managed Patients	No significant correlation between min. DS-CSA and leg or low back pain.
Lau 201725	70 Secondary Care Outpatients	No significant correlation between min. DS-CSA and leg pain when supine. On standing MRI a small negative correlation between min. DS-CSA and leg pain was identified (r = -0.20, p < 0.05).

Table B.24: Results relating to other measurements of disability and quality of life.

Outcome	Reference	Study Population	Results
JOA	Kanno 2012	88 Referrals for Surgical Management	A small positive correlation was identified between min. DS-CSA and JOA score (r = 0.27, p < 0.05) on standing MRI. On axial loading, the strength of correlation increased (r = 0.45, p < 0.001).
	Prasad 2016	48 Surgically Managed Patients	No significant correlation between min. DS-CSA and JOA score.
ZCQ	Burgstaller 2016	150 Secondary Care Outpatients	No significant correlation between min. DS-CSA or per segment DS-CSA and the symptom severity domain of the ZCQ.
	Marawar 2016	30 Secondary Care Outpatients	Positive correlation between ZCQ symptom severity domain score and DS-CSA at L2-L3 (p value not stated). No significant correlation found at any other level.
NCO	Weiner 2007	27 Secondary Care Outpatients	No significant correlation between NCOS and min. DS-CSA.
EQ-5D	Sigmundsson 2011	109 Surgically Managed Patients	No significant correlation between min. DS-CSA or the number of stenotic levels (threshold DS-CSA < 70 mm^2) and EQ-5D scores. No significant difference in EQ-5D scores between multi-level and single level stenosis groups.
BDI, BAI, PSQI	Hong 2015	73 Referrals for Epidural Adhesiolysis	No significant correlation between min. DS-CSA or number of stenotic levels (DS-CSA < 100 mm^2) and BDI, BAI or PSQI. No significant difference in BDI, BAI or PSQI between single and multi-level stenosis groups.
COMI	Mannion 2017	157 Surgically Managed Patients	No significant correlation between min. DS-CSA and COMI.

Appendix C

Study Protocols

Note: the following protocols are lightly edited versions of the protocols established at the beginning of the PhD and as such are written as though participant recruitment and data collection is a future event.

C.1 BOOST RCT imaging standard operating procedure

C.1.1 Purpose of BOOST imaging SOP

This SOP sets out the protocol for collection, processing and reporting of MRI data for participants of the BOOST randomised controlled trial. A standard operating procedure is required for compliance with good clinical practice and to standardise handling of significant unexpected findings in the MRI data that require urgent referral back to an appropriate clinician.

MRI data will be collected for all participants, except where contraindicated or where the patient declines consent. Imaging data is stored in a standardised format – digital imaging and communications in medicine (DICOM). This contains the image data itself and a "header" which contains patient information and MRI acquisition parameters for each study. DICOM files are stored within a picture archiving and communication system (PACS) maintained by each NHS trust. An electronic network exists for transfer of DICOM files between the PACS run by different trusts – the image exchange portal (IEP).

Participants will be referred for an MRI study of the lumbar spine if they have not already undergone one in the 12 months prior to randomisation. Where a recent clinical MRI study exists, it will be used in place of a specific research study. While the trial aims to exclude individuals with symptoms or signs of serious pathology, such as cauda equina syndrome or malignancy, it is possible that a clinical or research study will demonstrate findings that require urgent referral back to a patients GP or spinal surgeon for further management. It is expected that for studies performed for clinical reasons the performing NHS department will have already reported and acted upon the study results, however this will not be the case for studies performed for research purposes. Timely review and reporting of all research studies by the trial radiologist is therefore required, with appropriate action taken on any unexpected findings.

C.1.2 Definitions

- Clinical MRI — A magnetic resonance imaging study performed by an NHS trust for clinical purposes, as opposed to a study requested specifically for the purposes of the trial.
- DICOM — The file format for storage of MRI data.
- DICOM Header — The part of the DICOM file used for storage of patient demographic and study parameter information.
- image exchange portal (IEP) — a secure electronic system for transfer of DICOM files between NHS trusts
- Insignia — DICOM viewing software used by University Hospital Coventry and Warwickshire (UHCW)
- OSIRIX — DICOM viewing software for the MacOS operating system.
- picture archiving and communication system (PACS) — The electronic storage system for DICOM files maintained by the radiology department at UHCW.
- PACS team — The team of individuals responsible for operation and maintenance of the PACS and IEP systems.
- Research MRI — A magnetic resonance imaging study requested and performed specifically for the purpose of the trial.
- Warwick team — the University of Warwick based team coordinating MRI data collection and reporting, consisting of the trial radiology registrar and consultant, and based at UHCW.

C.1.3 Responsibilities

Trial radiology registrar—Dr Richard Gagen

A GMC registered medical practitioner holding a specialist training number in clinical radiology.

Responsibilities:

- Organisation of the transfer and collection of the MRI data once informed of participant information by the trial management team.
- Maintenance of a database of the status of all MRI data.
- Initial reporting of all collected studies.
- Notification of an appropriate responsible clinician (generally the patients general practitioner) in the event of a significant unexpected finding on a trial MRI study.

Trial radiology consultant—Prof. Charles Hutchinson

A GMC registered medical practitioner practising as a NHS consultant radiologist with special interest in musculoskeletal imaging.

Responsibilities:

- General clinical supervision of the trial radiology registrar.
- Review of all research MRI studies.
- Review of all MRI studies where there is a significant unexpected finding.

BOOST trial management team

The management team running the BOOST randomised controlled trial.

Responsibilities:

- Forwarding of MRI scan forms completed by recruiting centres to the trial radiology registrar.
- Managing storage of the backup of MRI data and reports taken to Oxford on a 3 monthly basis.
- Provision of the contact details of a participant's GP to the trial radiology registrar in the event of a significant unexpected clinical finding.

C.1.4 Specific procedure

MRI data collection

Notification of trial radiology registrar of a new participant. As part of baseline data collection for each participant the recruiting trial centre completes an MRI scan form. This records basic patient identity information, the location and date of any lumbar MRI study in the 12 months prior to randomisation and, if appropriate, whether they consent to have a MRI for the purposes of the trial. On completion, this form is sent to the BOOST management team, who forward it alongside an address for each participant, to the trial radiology registrar. The transfer is performed email link to a secure server where the MRI scan forms are temporarily stored.

Patients with a recent NHS clinical MRI of the lumbar spine. For patients with a clinical MRI of the lumbar spine performed at an NHS trust within 12 months of randomisation, the trial radiology registrar will complete a PACS transfer request form, using the participant information provided on the MRI scan form, and send this to the PACS team at UHCW. The information required by the PACS team to process the request includes: participant name, date of birth, address NHS number, gender, and the name of the hospital performing the study. The PACS team will contact the relevant NHS trust and request electronic transfer of the study to the UHCW PACS, via the IEP network, where it will reside for a period of 14 days before automatic deletion. During this period, the trial radiology registrar will export the images in DICOM format to an SSD using the Insignia software package. The DICOM images are automatically anonymized during the export process – with replacement of the DICOM header fields for the patient's name and ID with "Pat1" and a pseudo-random date of birth and study date. The DICOM files will be stored on the solid state drive (SSD) with a separate folder for each participant. Each folder will be named with the participant's trial ID. After transfer to the SSD the DICOM headers will be modified, using the DICOM browser software package, such that the name and ID fields are the participants trial ID.

Patients without a recent clinical MRI of the lumbar spine. For participants without a clinical MRI of the lumbar spine within 12 months of randomisation, the recruiting centre will request a research MRI study with the agreed local radiology provider. After receipt of the MRI scan form, the trial radiology registrar will contact the local radiology provider within a one-month period to check the date the participant has been booked for their MRI. After this date has passed the trial radiology registrar will confirm completion of the study. Where an NHS trust has performed the research MRI the procedure for transfer of the images will be as stated below. Where a private centre outside the IEP network has performed the research MRI the private centre will send the study via compact disc (CD) to the trial radiology team. On receipt of the CD the images will be exported to an SSD and anonymized and storied as below.

Data storage and backup. An electronic copy of all received MRI scan forms and all PACS transfer request forms produced will be kept on a password protected user account maintained by the UHCW trust. A hard-copy of all MRI scan forms received will be printed and stored with the SSDs in the locked office. On successful receipt of the images for each participant the electronic copy of the MRI scan form and PACS transfer request form will be deleted.

All MRI DICOM files will be stored in anonymized format upon a SSD kept within a locked office in the clinical sciences building on the UHCW campus. A backup of all MRI data will be made at the end of each week to a second backup SSD kept within the same office. At least every third month a copy of all data will be taken to Oxford to be stored on a secure server maintained by Oxford University. Any CDs received from private imaging centres will be stored in a locked office in the clinical sciences building until the completion of the trial.

Database of participant MRI studies. An excel spread sheet will be maintained by the trial radiology

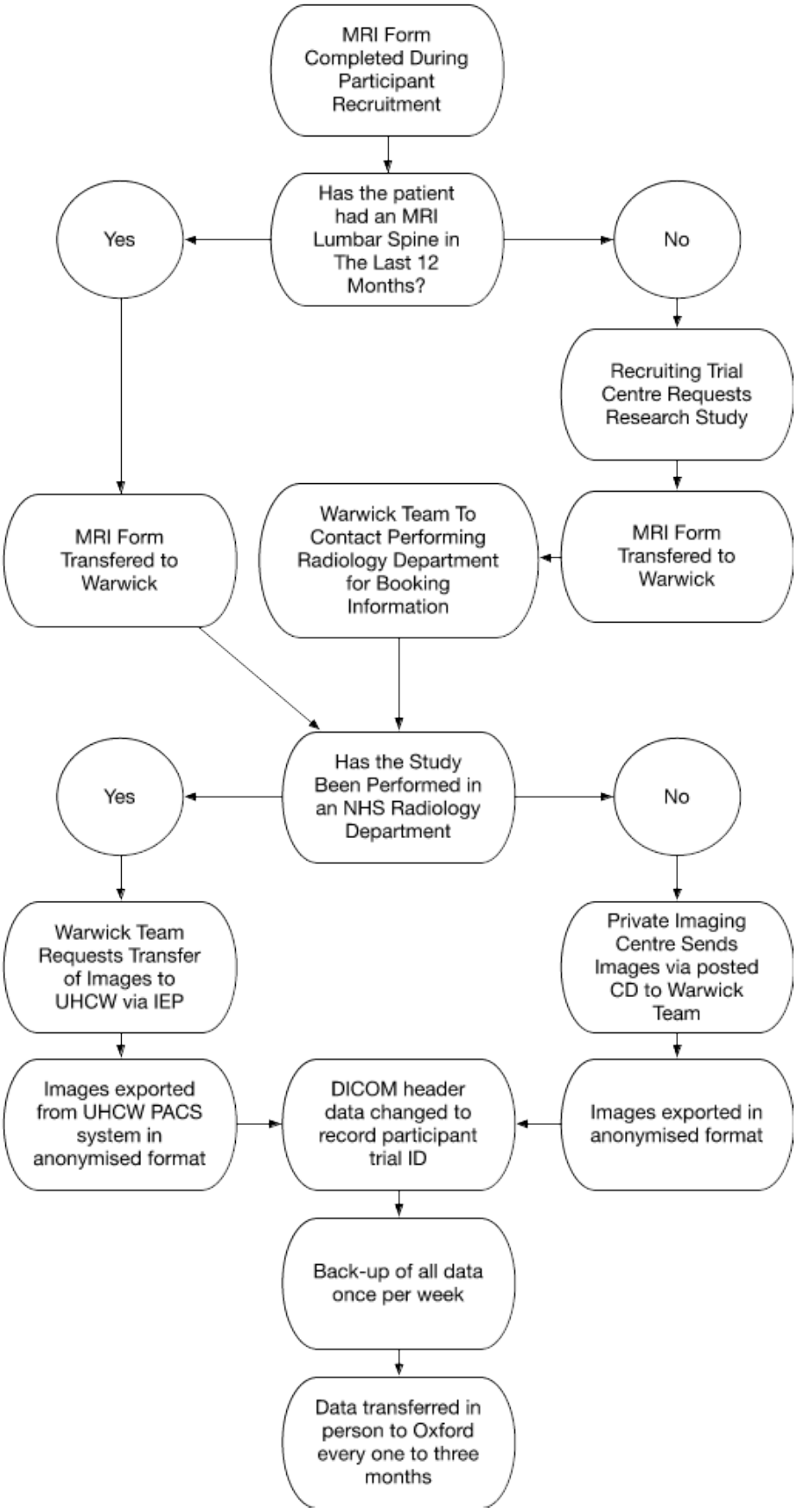


Figure C.1: Flowchart describing BOOST MRI data collection.

registrar containing anonymized data on the status of each participants MRI data. The date that each MRI scan form is received by the trial radiology registrar, date of study, image transfer request and receipt and reporting will be recorded. Once per month this data will be forwarded to the BOOST trial management team by email.

MRI reporting

Registrar radiology report. For all MRI studies received, the trial radiology registrar will complete a report, detailing any visualised pathology. The report will follow the format standard of radiology reporting in the United Kingdom and will identify the reported study by participant trial acID number and gender only. No specific quantitative metrics or qualitative grading systems will be used in the report, except where the grading system is likely to be understood by a spinal surgeon and will convey information that may change a participant's clinical management.

The trial radiology registrar will aim to complete the report within one week of receipt of images. During the peak of participant recruitment this may become difficult due to the large number of expected studies. In the event of a reporting backlog developing reporting of research MRIs will be prioritised as it is expected that the performing radiology department will have reported clinical MRIs.

Consultant radiology review. The trial radiology consultant will review the imaging and registrar report for all research MRI studies. Where there is discrepancy between the consultant's and registrar's interpretation of the imaging, the report will be altered to reflect the consultant's interpretation. Consultant review of research MRIs will take place within one week of the initial registrar radiology report.

Clinical MRIs will not be routinely reviewed by the trial radiology consultant, as it is expected that these studies will have been reported by a radiology consultant at the performing radiology department. In cases where the radiology registrar has uncertainty with regards to the interpretation of a clinical MRI, however, a consultant review of the study will be sought.

Reporting environment. DICOM images will be viewed for reporting using the software package Osirix, run on a mid-2015 27" iMac. The iMac is kept in the same office as the SSDs noted above. Reports will be produced in Microsoft Word. On completion, they will be stored as PDFs on the SSD hard drive. Backups will be made once per week to the backup SSD. At least once aver 3rd month a copy of all reports will be taken to Oxford for secure storage on a sever maintained by the University of Oxford.

Completion of trial. On completion of the BOOST trial participants who have undergone a clinical MRI will have the report of their MRI sent to them by the trial management team.

C.1.5 Managing significant unexpected findings

Definition. By the nature of the participant's age and presence of neurogenic claudication, it is expected that the MRI studies will show degenerative changes within the lumbar spine. These degenerative changes will include facet joint and ligamentum flavum hypertrophy, intervertebral disc bulge, protrusion or extrusion, and/or spondylolisthesis. It is expected that in most participants the degenerative changes will result in stenosis of the central spinal canal, lateral recesses or neural exit foramina.

Pathology outside the above, that requires urgent clinical assessment and management, is for the purposes of this document defined as a significant unexpected finding. This may include, but is not limited to:

- Large disc protrusion/extrusion compressing the cauda equine
- Suspected neoplastic disease
- Acute fracture
- Abdominal aortic aneurysm

Identification of significant unexpected findings. Where a significant unexpected finding is identified on initial reporting by the trial radiology registrar, corroboration of the nature of the unexpected pathology and advice will be sought from the consultant radiologist. A consultant agreed report will be produced with clear statement of the identified significant pathology. The report will identify the participant by trial ID, name, date of birth and NHS number as recorded on the MRI scan form. The report will contain contact details for the trial radiologists and information on how to access the DICOM files in the event of admission of the participant to a secondary care centre.

Informing the participants GP. After completion of the consultant agreed report and comparison with any pre-existing report for clinical MRIs, a decision will be made by the trial radiology consultant, with regards to whether the participants GP needs to be informed. This decision will be made based upon the nature of the identified pathology and the likelihood of harm to the participant if not acted upon.

Where a decision is taken to inform the participants GP, the trial radiology registrar will contact the BOOST trial management team by telephone, who will obtain details of the participants GP and pass these by phone to the trial radiology registrar. In normal working hours, the trial registrar will phone the relevant GP surgery and speak to the duty doctor to inform them of the finding. If out of hours and the finding is of significant urgency that it would likely lead to admission overnight, the trial radiology registrar will contact the out of hours GP service in the relevant area to inform them of the finding. In all cases the radiology report will be faxed to the GPs surgery.

Availability of research MRI studies for secondary care review. In the event of a significant unexpected

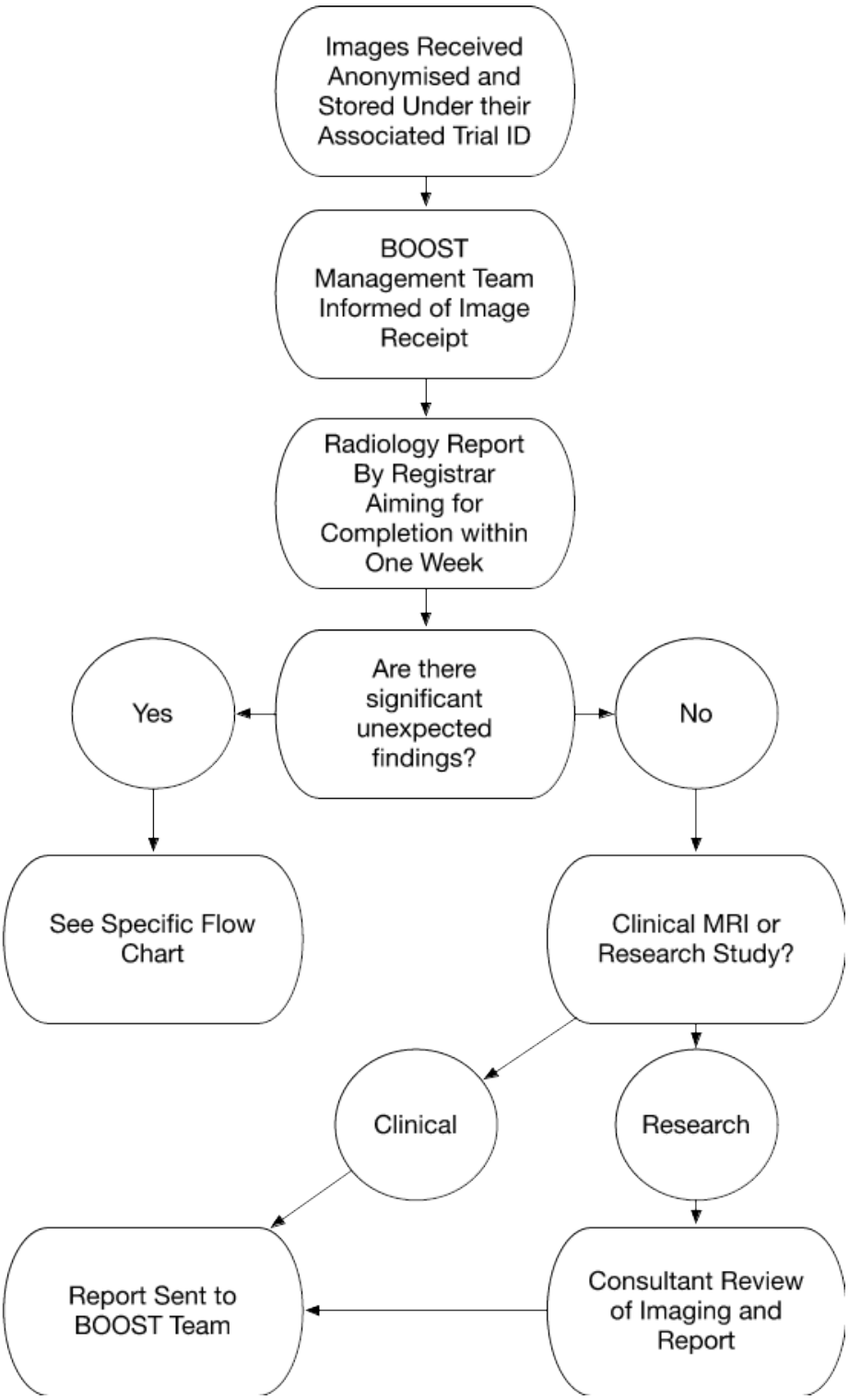


Figure C.2: Flowchart describing the procedure for the clinical reporting of BOOST MRI studies.

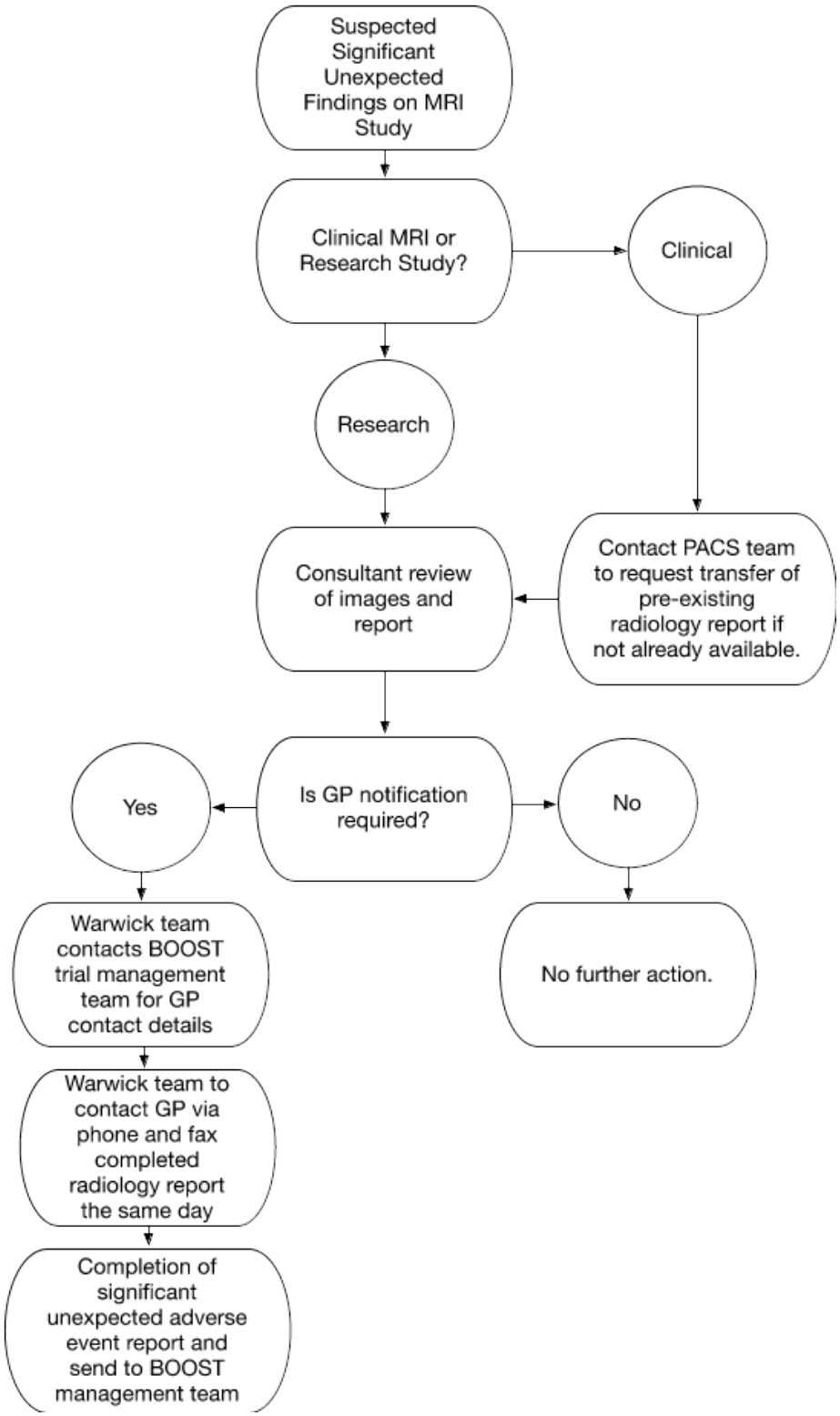


Figure C.3: Flowchart describing the management of significant unexpected findings on a BOOST MRI study.

finding and the admission of said patient to a secondary care centre, it is highly likely that the treating physicians and associated radiology team will wish to review the study. Where the study has been performed in an acNHS radiology department the secondary care centre will be able to transfer the images to their PACS via IEP. For studies performed privately and transferred to the trial radiologists by CD, the study, on discovery of the significant unexpected finding, will be added to the UHCW PACS system under the participant's details, allowing the secondary care centre to transfer the study to them via IEP.

Completion of significant adverse event report. For any significant unexpected pathology identified, where a decision has been made to inform the participants GP, a significant adverse event report will be completed and forwarded to the BOOST trial management team.

C.2 Imaging control group recruitment

C.2.1 Study design overview

The following protocol sets out the process for recruitment of an imaging control group for comparison with participants of the BOOST randomised controlled trial. BOOST participants are recruited based on the presence of neurogenic claudication defined for the purposes of the trial by the following questions:

1. In the past 6 WEEKS, have you had back pain and/or pain or other symptoms such as tingling, numbness or heaviness that travelled from your back into your buttocks or legs?

2. Does standing make the pain or symptoms in your buttocks or legs worse?

3. Does walking make the pain or symptoms in your buttocks or legs worse?

4. Does sitting down make the pain or symptoms in your buttocks or legs better?

5. Does bending forward (for example to push a shopping trolley) make the pain or symptoms in your buttocks or legs better?

Patients who answer the first question as yes, and at least one of questions 2 to 5 as yes are judged to have neurogenic claudication.

This study aims to recruit a population of patients who have been referred by their GP for an MRI scan for LBP, but who do not have neurogenic claudication. The MRI scans collected from these participants will be used as a comparison group with the MRIs collected from the BOOST participants. It is hoped that by comparing the various MRI measurements that have been used to characterise spinal stenosis between two groups, the measurement that best separates them and its optimum threshold for doing this can be established.

The full clinical importance of this measurement and associated threshold would then be established by further work, both within the BOOST trial and potentially further studies, but could potentially act as a diagnostic threshold for lumbar spinal stenosis.

C.2.2 Objective and outcome measures

Research objective

To assess whether measurements of the spinal canal on magnetic resonance imaging of the lumbar spine can differentiate reliably between people with neurogenic claudication and those with non-specific lower-back pain, what measurement does this most reliably and with what diagnostic threshold.

Primary hypothesis

The dural-sac cross-sectional area at the site of most severe stenosis is smaller in patients with neurogenic claudication compared to LBP groups and can separate the two groups with an acceptable sensitivity and specificity.

Secondary hypothesis

As for the primary hypothesis but with the following anatomical measurements:

- Quantitative measurement of:
  - Lateral recess height
  - Neural exit foramen height
- Qualitative measurement of:
  - Relationship between CSF and Cauda Equina as per Schizas et al. (2010).
  - Compromise of the central canal as per Lurie et al. (2008a).
  - Foraminal nerve-root impingement as per Pfirrmann et al. (2004).

Measurements relating to both the primary and secondary hypothesis listed above will be measured at each spinal level for which axial imaging is available. They have been chosen based upon a recently published consensus criteria for radiology reporting in spinal stenosis (Andreisek et al. 2014).

C.2.3 LBP study population

All patients over 65 referred to the University Hospitals Coventry and Warwickshire (UHCW) for MRI of the lumbar spine by their general practitioner with a

primary history of lower back pain will be considered for recruitment.

With the exception that patients with neurogenic claudication are excluded rather than included, the study inclusion and exclusion criteria are listed below and have been chosen to match as closely as possible the BOOST trial inclusion and exclusion criteria to ensure comparability between BOOST and control groups.

Inclusion criteria

- Registered with a primary care practice
- Referred by their GP for MRI examination at UHCW of the lumbar spine for investigation of low back pain.
- Aged 65 years and over.
- Participant is willing and able to give informed consent for participation in the trial.
- Patient does not report symptoms consistent with neurogenic claudication.

Exclusion criteria

- Living in a residential care or nursing home.
- Has a terminal condition with a life expectancy of fewer than 6 months.
- Unable to walk 3 meters without the help of another person.
- On a surgical waiting list for back pain or related conditions.
- Cognitive impairment (defined as an Abbreviated Mental Test score of 6 or less).
- Registered blind.
- Unable to follow verbal instructions.
- Signs of serious pathology on MRI lumbar spine that requires immediate referral for further management (e.g. neoplasia, signs of acute cauda equina compression)
- Under the age of 65

C.2.4 Study procedures

Recruitment of LBP participants

The study will take place at a single centre (UHCW). Potential participants will be identified by review of GP referrals for lumbar spine MRI by a member of the clinical team. Such participants will be sent a study invitation letter alongside their MRI booking information by post. This is usually done up to a month in advance of the planned MRI study date.

On arrival to the department, patients with appointments for an MRI scan are booked in at reception and undertake an MRI safety questionnaire with either a radiographer or radiology department assistant. At this point, potential participants will be asked whether they would consider allowing the MRI images taken to be used for research purposes. If they indicate they would consider this, a participant information sheet will be given, and a member of the study team will attend to give further information and check eligibility criteria.

Potential participants will be asked the first neurogenic claudication self-report question from the BOOST trial, detailed above. Only those who answer no will undergo further screening. Such patients, who do not have pain radiating into the buttocks or legs, are highly unlikely to have neurogenic claudication and are potentially suitable for the study. The remaining inclusion and exclusion criteria will then be checked and patients who meet these criteria will then be asked to provide informed consent and recruited into the study.

Informed consent

Written informed consent will be taken by a member of the study team. Before taking consent, time will be given to read through the participant information sheet.

The consent process will take place before the MRI study and will be conducted in the designated area within the MRI department for consent and procedure preparation. This area provides sufficient privacy for this purpose. The consent process will involve both an explanation of the study, its use of the patients MRI images and data and the fact it will not change the patient's subsequent management. An opportunity to ask questions will be given. The process is expected to take approximately 10 to 15 minutes, to allow time to properly explain the study and answer any questions the participant may have.

We do not seek to include participants from vulnerable groups and will exclude participants who are unable to provide consent themselves due to cognitive or communication issues. If a potential participant is unable to fully understand the verbal consent process due to language differences, an attempt will be made to find an interpreter. Where no interpreter can be found in a reasonable time frame the patient will be excluded from the study. Family members will not be used for translation due to the difficulties posed in ensuring a fair translation free of bias.

MRI study

The MRI study will be performed by a MR qualified radiographer and takes approximately 20 minutes to perform. The standard MRI protocol for GP low back pain referrals at UHCW includes both sagittal T1 and T2 weighted images of the lumbar spine from T8 to S2 and axial T2 weighted images of at least the lower 3 lumbar discs, with additional levels imaged at the discretion of the performing radiographer.

At UHCW the axial imaging is most often performed as a single block, however, for the purposes of the trial these will be performed angled to match the intervertebral disc. This change is made to allow easier comparison to other centres where it is the most common method used. It does not significantly affect scan time or study interpretation. No other changes to the standard MRI protocol is planned.

The study does not involve exposure to ionising radiation or administration of intra-venous contrast agent.

Data handling

On completion of the participants MRI the study images will be exported in anonymized format for storage and subsequent analysis.

All participant's MRI scans will have a clinical report produced. Any patients identified to have serious pathology requiring urgent management (e.g. neoplasia or signs of acute cauda equina compression) will be excluded at this stage. The clinical report will not be used for research purposes and will be distributed to the patient's general practitioner following the general radiology reporting protocol established at UHCW, as per normal radiology department practice.

In addition to the MRI, data will be collected on age, gender, BMI and ethnicity. This data will be used to ensure the sample is comparable to the BOOST participants in these respects.

Withdrawal of participants

As no intervention or change in management is involved, participants who do agree on the day but subsequently change their mind, can effectively fully withdraw from the study up to the point of anonymisation before its publication. Notification of withdrawal will be either verbal on the day of the study or via written communication, by post or email address provided on the participant information sheet.

End of study definition

The study will be complete after recruitment and MRI scan of the last study participant.

C.2.5 Statistics and analysis

Blinding of MRI observers to participant group

Each MRI study will be assigned and identified only by a random identification number, separate to the participant identification number. The random identification number will be generated by an automated algorithm run by a member of the study team not involved in data collection. That individual will keep a secure record of the key linking study ID to participant ID and this will not be available to other members of the research group until completion of data collection. The order that MRIs studies are analysed in will be randomised to prevent guessing of participant group by the order of MRIs for analysis.

Randomisation of MRI study ID will be performed in blocks to allow processing of the MRI studies to begin before all participant MRIs have been performed and collected.

Data analysis plan

All MRI studies will be assessed by at least a single observer. For each participant, the set of measurements of the lumbar canal detailed in section 7 will be recorded at all levels for which axial images are available. The primary observer will re-measure at least 20% of the studies and a second observer will measure at least 20% of the studies. Both observers will be trained using a handbook giving details on each measurement alongside examples. Intra- and inter-rater reliability will be assessed by Cohen's Kappa statistic for categorical variables (Cohen 1960) or the intra-class correlation for continuous variables (Bartko 1976). In the event of unacceptable intra- and inter-rater reliability, as determined by discussion among the study group, the number of repeat measurements will be increased, with averages for each measurement used in subsequent data analysis.

Data analysis will be performed primarily by ROC curve analysis (Obuchowski 2005). The AUC for each ROC curve will be calculated as a summary of the overall diagnostic performance of each measurement and presented alongside 95% confidence intervals. In the event the ROC curves of the different measurements cross, a partial AOC analysis prioritising specificity will be considered, after discussion between the study team. The optimal threshold for separating the two participant groups for promising measurements will be calculated using the Youden index (Fluss et al. 2005) followed by calculation of the sensitivity and specificity of the optimal threshold.

The computer science department at Warwick has produced an algorithm that applies neural network based landmark recognition to MRI spine data. This process can extract morphological information from the MRIs and then can be trained to automatically grade stenosis and potentially differentiate normal from abnormal studies. The computer software to apply this has already been developed. Training and testing this system for use in spinal stenosis is a potential use of the data this study will produce and will be explored after completion of the primary data analysis. All data used for this analysis will be fully anonymized.

Sample size

As of January 2018, BOOST has recruited 269 participants with MRIs available for 159 (59%). As additional MRIs are performed for participants without scans this percentage is expected to rise to approximately 80%. BOOST plans to recruit a total of 402 participants and therefore we expect to have around 320 MRI s from BOOST participants.

Using an allocation ratio of 0.5 (2 BOOST participants to 1 low back pain participant) and an alpha of 0.05, a sample size of 200 BOOST participants and 100 low back pain participants would give a 90% power to detect an area under the curve (AOC) of 0.6 (Obuchowski 2005). If the allocation ratio is reduced to 0.35 then a sample size of 258 BOOST participants

and 90 low back pain participants would be required to give the same power.

In total we plan to recruit 150 participants. This should give adequate power even in the event of patients withdrawing consent or under-performance of the BOOST, both in terms of recruitment or provision of MRI scans. A larger sample will also provide more studies for training and testing the machine learning system, which requires a large sample that can be split into both a training group, and a group that the resultant model will be tested against.

UHCW performs approximately 87 GP MRI spines per month (data Jan to May 2017). Assuming a 60% eligibility and 60% consent rate for eligible individuals, recruitment will take approximately 4 months.

C.2.6 Ethical approval, governance and risk assessment.

Ethical approval and research governance

The study will be conducted in compliance with the principles of the ICH GCP guidelines and in accordance with all applicable regulatory guidance, including, but not limited to, the UK Policy Framework for Health and Social Care Research. Ethical approval for this study will be sought from the Research Ethics Committee combined with Health Research Authority (HRA) approval via the IRAS electronic application system.

No study activities will commence until favourable ethical opinion and HRA approval has been obtained. Progress reports and a final report at the conclusion of the trial will be submitted to the approving REC within the timelines defined by the committee.

The UHCW research and development department will be asked for confirmation of capacity and capability before commencement of the study.

Audit and inspection

The study may be monitored by the Research and Development Department at UHCW as representatives of the Sponsor, to ensure that the study is being conducted as per protocol, adhering to Research Governance and GCP. The approach to, and extent of, monitoring will be specified in a trial monitoring plan determined by the risk assessment undertaken prior to the start of the study.

Public and patient involvement

Patient group consultation was undertaken in the design of the BOOST trial. As this observational study is an addition to the BOOST trial and does not involve an intervention no further patient group consultation has been undertaken.

Data protection

The study will comply with the Data Protection Act 1998. Personally identifiable data of participants will be stored on paper in a locked office. This will include the consent form, eligibility screening forms and a table of patient names, date of birth, address and NHS numbers, linked to a individual study specific identification number assigned upon recruitment. A copy of this table will be stored electronically on the NHS computer system, in encrypted format and accessible only to the study team. The study ID number will be used for all other documentation and data analysis.

The MRI images will be exported from the UHCW PACS (picture archive and communication system), the electronic system which stores all the images from all radiology investigations, to a solid state hard drive for transfer to a university computer located in a locked office in the CSB building on the UHCW site. At the point of export from PACS the MRI images are fully anonymized and will be identified only by the participant ID number.

Only members of the study team and authorised representatives of the sponsor will have access to study data.

Risk assessment

As a cross-sectional observational study recruiting participants already planned for an MRI study, we do not anticipate any significant risk to participants outside that of their normal care.

Potential risks and mitigation:

- Potential harm may arise from exposure to the magnetic field gradients if a study is performed in an individual with an implanted medical device (dysfunction of the device), with incorrect patient positioning (burns) or if a metallic object is brought into the study room and acts as a projectile. All patients are planned for a MRI as part of their routine care and so no additional risk is required for the trial. The study will be performed by a qualified radiographer following the normal safety procedures at UHCW.
- For the purposes of the study the orientation of one imaging sequence will be changed to standardise with other NHS trusts. This potentially could reduce the diagnostic performance of the study due to the performing radiographer and interpreting radiologist being more familiar with the orientation standard to UHCW. In mitigation, where possible the studies on potential research participants will be performed by a research radiographer who is familiar with this protocol. The proposed orientation is the standard at most other NHS trusts and all radiologists will have some experience in its use.
- Potential loss of confidential information. In mitigation data will be stored as per the data collection and storage section and with respect to the relevant legislation.

C.2.7 Study dissemination

On completion of the study it is intended to publish the study within a peer reviewed journal. A plain text summary of the study results will be posted to participants who have given their permission for this.

C.3 BOOST outcome measures

Primary:

- Oswestry disability index (Fairbank et al. 2000)

Secondary:

- Back and leg symptoms:
  - Back pain troublesomeness (Parsons et al. 2006)
  - Swiss Spinal Stenosis Questionnaire (symptom subscale) (Stucki et al. 1996)
- Quality of life:
  - EQ5D-5L (EuroQol Group 1990)
- Frailty:
  - Tilburg Frailty Index (Gobbens et al. 2010)
  - Hand grip strength (H. C. Roberts et al. 2012)
- Physical activity (Rapid Assessment Disuse Index (Shuval et al. 2014)):
  - Time spent moving around on feet
  - Time spent sitting
- Mobility and Balance:
  - Change in mobility over the past 6/12 months (baseline/followup respectively).
  - 6 minute walk test (Bennell et al. 2011)
  - Prevention of falls network Europe self-report of falls and falls related injury (Lamb et al. 2005)
  - Short physical performance battery (Guralnik et al. 1994)
- Self-efficacy:
  - Confidence in walking 0.5 miles (Newell et al. 2012)
  - Self-efficacy recovery and maintenance related to performing home activities (Luszczynska et al. 2011)
- Exercise adherence:
  - Self-report of adherence to home exercise program.
- Habit (automaticity):
  - Self-report habit index (B. Gardner et al. 2012)
- Fear avoidance:
  - Fear avoidance beliefs questionnaire (Waddell et al. 1993)
- Global rating of perceived change:
  - Change in back and leg problems (Kamper et al. 2009)
- Satisfaction in:
  - Provided exercises
  - Change in back and leg problems
  - Increase in physical activity
- Spinal alignment:
  - C7 to wall measure (Wiyanad et al. 2018)
- Health resource use:
  - Self-report of medication use
  - Client Service Receipt Inventory (Thornicroft et al. 2001)

The degree of thoracic kyphosis is measured using a modification of the occiput to wall distance measurement. (Wiyanad et al. 2018) The participant stands with their sacrum and back against a wall and the distance from the C7 spinous process to the wall is measured using a ruler. This method aims to improve the reliability of occiput-wall measurement by removing variability caused by head movement.

The SPPB measures three aspects of physical performance, each rated on a scale of 0–4. Walking speed was measured on a eight feet long course at the participants usual walking speed with the faster of two attempts recorded. Standing balance was judged on the participants ability to maintain three test positions for ten seconds. Participants were finally scored on the time it took them to perform 5 transitions to a standing positron from sitting with arms folded across the chest in a straight-backed chair. The scores from all three tests are added to provide the final SPPB score (Guralnik et al. 2000).

The 6-MWT measures the distance a participant is able to walk in 6 minutes around a indoor straight and flat course with a minimum straight length of 10 meters. (Bennell et al. 2011) The participant was also asked to indicate when they began to experience symptoms during the test so the distance could be recorded.

BOOST baseline data collection

Demographic information:

- Age and sex
- Current alcohol intake and smoking status
- Ethnicity
- Relationship status
- Postcode
- Type of housing
- Current occupation
- Education
- Unpaid/paid care status
- Household income
- Height
- Weight

Health Status:

- Participants perceived ability to manage their condition
- Self-report of current health conditions
- Other pain problems: Nordic pain questionnaire (Kuorinka et al. 1987)
- Attitudes to Ageing Questionnaire (Laidlaw et al. 2007)
- Risk of developing persistent disabling back related symptoms: STarT Back Screening Questionnaire (Hill et al. 2008)

Mobility:

- Use of walking aids
- Self-rated walking speed

Psychological factors:

- Self-efficacy for exercise (short version) (Resnick et al. 2000)
- Intention to carry out home exercise program

Imaging:

- MRI lumbar spine (see § 3.2.8)

Appendix D

Code

The following code was used to model nerve root compression within the lumbar spine and was described in full in §8.4. It was coded for Python 3.8.5 (Built October 24 2020).

The software in the following section is provided "as is", without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall the author or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings in the software.

```
## Imports
from enum import Enum, IntEnum
import os
import re
import pandas as pd
import numpy as np
from datetime import datetime
from progressbar import progressbar


## Enums
class Side(Enum):
    """Side of the body."""

    LEFT = 1
    RIGHT = 2

class NLevel(IntEnum):
    """All nerve roots simulated within the model."""

    L1 = 1
    L2 = 2
    L3 = 3
    L4 = 4
    L5 = 5
    S1 = 6
    S2 = 7

class VLevel(IntEnum):
    """All vertebral levels at which potential nerve root compression is modelled"""

    L1 = 1
    L2 = 2
    L3 = 3
    L4 = 4
    L5 = 5

class CS(Enum):
    """Sites at which a nerve root may be compressed within a single spinal level.

    CENTRAL refers to the central canal
    LATERAL refers to the lateral recess
    FORAMINAL refers to the neural exit foramen
    NONE is used to represent spinal levels were a particular nerve root cannot
      be compressed"""

    CENTRAL = 1
    LATERAL = 2
    FORAMINAL = 3
    NONE = 4
```

```
class CombnRule(Enum):
    """Rule for combining compression information from quantitative and
    qualitative grading systems to judge whether nerve root compression is
    occurring at a single potential compression site.

    AND - threshold for both quantitative and qualitative measurement must be met
    OR - threshold for either the quantitative or qualitative measurement must be met
    QUANT - only quantitative data is considered
    QUAL - only qualitative data is considered"""

    AND = 1
    OR = 2
    QUANT = 3
    QUAL = 4


class NerveRoot():
    """Class which stores data and functions relating to a single nerve root.
    Each spine consists of 7 paired nerve roots L1 through S2."""

    def __init__(self, side, level, measurements):
        """Create a new NerveRoot.

        SIDE - laterality of the nerve root.
        LEVEL - neural exit foramen through which the nerve root exits the spine
        MEASUREMENTS - reference to the measurements of canal size within which
        the NerveRoot exists"""

        ## Set level and side
        self.side = side
        self.level = level
        ## Get possible compression sites
        self.poss_cs = self._set_possible_cs(level)
        ## Extract local canal size and qualitative judgement of compression for
        ## each VLevel
        self.quant_diameters = measurements.get_quant(self.poss_cs, side)
        self.qual_diameters = measurements.get_qual(self.poss_cs, side)

    def _set_possible_cs(self, nlevel):
        """Returns a dictionary which contains the location of a given nerve
        root for all vertebral levels within the lumbar canal - i.e. the L5
        nerve root is within the central canal at L3 and lateral recess at L4

        Output dictionary structure - {VLevel : CS}"""

        possible_cs = dict()
        ## Iterate through all lumbar vertebral levels
        for vlevel in VLevel:
            if vlevel < nlevel - 1:
                ## 2 levels and above away from the neural roots exit the root
                ## is in the central canal
                possible_cs[vlevel] = CS.CENTRAL
            elif vlevel == (nlevel - 1):
                ## One level away from the nerve roots exit the root is in the
                ## lateral recess
                possible_cs[vlevel] = CS.LATERAL
            elif vlevel == nlevel:
                ## At the level of the nerve roots exit the root is within the
                ## neural exit foramen.
                possible_cs[vlevel] = CS.FORAMINAL
            else:
                ## Below the level of the nerve roots exit the root has left the
                ## canal.
                possible_cs[vlevel] = CS.NONE
        return possible_cs

    def debug_print(self):
        """Prints the relevant quantitative measurements canal size and
        qualitative judgement of nerve root impingement associated with the nerve
        root at each lumbar vertebral level."""

        print(f"Nerve root: {self.side.name} {self.level.name}")
        print("Quant: " + self._form_print_string(self.quant_diameters))
        print("Qual:  " + self._form_print_string(self.qual_diameters))

    def _form_print_string(self, canal_measurements, digits = 1):
        """Iterates through all possible compression sites for the nerve root
        forming a formatted string containing the measurments at those
```

```
        compression sites.

        CANAL_MEASUREMENTS - dictionary of measurements, quantitative or
        qualitative, around the nerve root at each VLevel.
        DIGITS - the number of digits to round to in the returned string.
        """

        print_string = ""
        ## Iterate through possible compression sites.
        for level, cs in self.poss_cs.items():
            ## Extract relevant measurement
            value = round(canal_measurements[level], digits)
            ## Add to formatted string to be returned.
            print_string = print_string + f"{level.name}({cs.name[0]}) {value} "
        return(print_string)

    def _process_thresholds(self, thresholds):
        """Extracts from a dictionary of thresholds the relevant threshold for
        each possible compression site. Returns separate dictionary for
        quantitative and qualitative measurements.

        Output dictionary structure = {VLevel : Measurement Threshold}

        THRESHOLDS - a dictionary containing thresholds for all spinal
        measurement types."""

        quant_thresholds = dict()
        qual_thresholds = dict()
        ## Iterate through compression sites.
        for level, poss_cs_site in self.poss_cs.items():
            ## If root in central canal add central measurement thresholds.
            if poss_cs_site == CS.CENTRAL:
                quant_thresholds[level] = thresholds["dscsa"]
                qual_thresholds[level] = thresholds["schizas"]
            ## If in lateral canal add lateral measurement thresholds.
            if poss_cs_site == CS.LATERAL:
                quant_thresholds[level] = thresholds["lrd"]
                qual_thresholds[level] = thresholds["lgrade"]
            ## If in foramina add foraminal measurement thresholds.
            if poss_cs_site == CS.FORAMINAL:
                quant_thresholds[level] = thresholds["nfd"]
                qual_thresholds[level] = thresholds["fgrade"]
            ## If root has left the canal add NaN place-holder
            if poss_cs_site == CS.NONE:
                quant_thresholds[level] = np.nan
                qual_thresholds[level] = np.nan
        return(quant_thresholds, qual_thresholds)

    def check_compression(self, thresholds, combn_rule, debug = False):
        """For a given set of measurements thresholds, returns a Float
        containing the number of sites at which the nerve root is compressed.

        THRESHOLDS - dictionary of measurement thresholds
        COMBN_RULE - CombnRule object specifying how quantitative and qualitative
        measurements are to be combined.
        DEBUG - if true print debug information"""

        ## Extract relevant thresholds for each possible compression site
        quant_thresholds, qual_thresholds = self._process_thresholds(thresholds)

        quant_compression = list()
        qual_compression = list()
        ## Iterate through vertebral levels buiding a list of whether each
        ## measurement is beyond the threshold for stenosis
        for level in VLevel:
            quant_compression.append(self.quant_diameters[level] <= \
                                     quant_thresholds[level])
            qual_compression.append(self.qual_diameters[level] >= \
                                    qual_thresholds[level])

        ## Combine quantitative and qualitative measurements
        if combn_rule == CombnRule.QUANT:
            compression = quant_compression
        elif combn_rule == CombnRule.QUAL:
            compression = qual_compression
        elif combn_rule == CombnRule.OR:
            compression = np.logical_or(quant_compression, qual_compression)
        elif combn_rule == CombnRule.AND:
```

```
            compression = np.logical_and(quant_compression, qual_compression)

        ## Sum sites of compression
        return_value = sum(compression)

        if debug:
            self.debug_print()
            print("Quant comp:  ", quant_compression)
            print("Qual comp:   ", qual_compression)
            print("Compression: ", compression, f"{return_value}\n")

        return(return_value)


class Measurements():
    """Stores and processes measurements associated with a single MRI scan."""

    def __init__(self, id_num, data):
        """Creates a new Measurements object from a given row of the MRI
        measurements data frame.

        ID_NUM - blinding id of MRI scan to be extracted
        DATA - wide data frame containing MRI data"""

        ## Extract row from data frame
        row = data[data["blindingID"] == id_num]
        if row.shape[0] != 1:
            raise RuntimeError("Incorrect row shape!")

        ## Extract dictionaries of measurements from the data frame row
        self.dscsa = self._extract_central("dscsa", row)
        self.schizas = self._extract_central("schizas", row)
        self.lrd = self._extract_non_central("lateralDepth", row)
        self.nfd = self._extract_non_central("foraminaDiameter", row)
        self.lgrade = self._extract_non_central("lateralGrade", row)
        self.fgrade = self._extract_non_central("foraminaLee", row)

        ## Convert qualitative grades strings to interger scale
        self.schizas = self._convert_grades(self.schizas, "schizas")
        self.lgrade = self._convert_grades(self.lgrade, "other")
        self.fgrade = self._convert_grades(self.fgrade, "other")

    def debug_print(self):
        """Print all measurements extracted and stored from data-frame"""

        print("DSCSA: " + self._form_debug_string(self.dscsa))
        print("Schizas: " + self._form_debug_string(self.schizas))
        print("Left LRD: " + self._form_debug_string(self.lrd[Side.LEFT]))
        print("Right LRD: " + self._form_debug_string(self.lrd[Side.RIGHT]))
        print("Left NFD: " + self._form_debug_string(self.nfd[Side.LEFT]))
        print("Right NFD: " + self._form_debug_string(self.nfd[Side.RIGHT]))
        print("Left LGrade: " + self._form_debug_string(self.lgrade[Side.LEFT]))
        print("Right LGrade: " + self._form_debug_string(self.lgrade[Side.RIGHT]))
        print("Left FGrade: " + self._form_debug_string(self.fgrade[Side.LEFT]))
        print("Right FGrade: " + self._form_debug_string(self.fgrade[Side.RIGHT]))

    def _form_debug_string(self, measurement_list, digits = 1):
        """Format and return a string of measurements for printing from a given
        measurement list.

        MEASUREMENT_LIST - a list of measurements in order of vertebral level"""

        print_string = ""
        for value in measurement_list.values():
            value = round(value, digits)
            print_string = print_string + f" {value}"
        return(print_string)

    def _extract_central(self, var_name, row):
        """Extract central measurement VAR_NAME from a single row from the MRI data
        frame. Return a dictionary linking vertebral level to the variable
        measurement.

        Output dictionary structure - {VLevel : measurement}

        VAR_NAME - the variable to extract
        ROW - a single row data frame containing the MRI measurements.
        """
```

```
        result = dict()
        ## Iterate through columns matching the variable name
        for label, series in row.filter(regex = var_name + "_L").iteritems():
            ## Extract the level from the column name
            level = re.search(var_name + "_(L\d)", label).group(1)
            ## Add extracted measurement to dictionary
            result[VLevel[level]] = series.tolist()[0]
        return(result)

    def _extract_non_central(self, var_name, row):
        """Extract lateralised measurement VAR_NAME from a single row from the
        MRI data frame. Return a dictionary linking vertebral level and side to
        the variable measurement.

        Output dictionary structure - {VLevel : measurement}

        VAR_NAME - the variable to extract
        ROW - a single row data frame containing the MRI measurements."""

        ## Initialise dictionary with a sub-dictionary for each side
        result = {Side.LEFT : dict(), Side.RIGHT : dict()}
        ## Regex for extracting side information
        side_rx = "[0-9A-Z]*_(Left|Right)"
        ## Iterate through columns matching the variable name
        for label, series in row.filter(regex = var_name + "_L").iteritems():
            ## Extract level and side information from column name
            match = re.search(var_name + "_(L\d)" + side_rx, label)
            level = VLevel[match.group(1)]
            side = Side[match.group(2).upper()]
            ## Assign measurement to dictionary
            result[side][level] = series.tolist()[0]
        return(result)

    def _convert_grades(self, input_grades, mode):
        """Convert extracted grade strings stored in a dictionary into an
        integer scale between 1 and 3.

        Output dictionary structure - { [Side] : VLevel : measurement }

        INPUT_GRADES - dictionary of qualitative grades, with vertebral
        levels/sides as keys
        MODE - either "schizas" or "other" to determine how to convert grades.
        """

        result = dict()
        if mode == "schizas":
            # Dictionary to convert grades to integers
            conversion_grades = {"Grade A" : 0, "Grade B" : 1, "Grade C" : 2, "Grade D" : 3}
            # Iterate through grades converting to the interger
            for level, value in input_grades.items():
                result[level] = conversion_grades.get(value, np.nan)
        elif mode == "other":
            # Dictionary to convert grades to integers
            conversion_grades = {"Grade 0" : 0, "Grade 1" : 1, "Grade 2" : 2, "Grade 3" : 3}
            # Iterate through sides
            for side, sided_grades in input_grades.items():
                result[side] = dict()
                # Iterate through grades converting to the integer
                for level, value in sided_grades.items():
                    result[side][level] = conversion_grades.get(value, np.nan)
        else:
            raise RuntimeError("Mode must be 'schizas' or 'other'.")
        return(result)

    def get_quant(self, poss_cs, side):
        """Given a dictionary of possible compression sites for a nerve root,
        return a dictionary containing the relevant quantitative measurement for
        each level

        Output dictionary structure - { VLevel : measurement }

        POSS_CS - dictionary of possible nerve root compression sites
        SIDE - CS.SIDE object indicating nerve root side
        """

        quant_values = dict()
        ## Iterate through possible compression sites and add the relevant
```

```
        ## measurement for each site to the dictionary.
        for level, cs in poss_cs.items():
            if cs == CS.NONE:
                quant_values[level] = np.nan
            elif cs == CS.CENTRAL:
                quant_values[level] = self.dscsa[level]
            elif cs == CS.LATERAL:
                quant_values[level] = self.lrd[side][level]
            elif cs == CS.FORAMINAL:
                quant_values[level] = self.nfd[side][level]
            else:
                raise RuntimeError("Invalid CS value.")
        return(quant_values)

    def get_qual(self, poss_cs, side):
        """Given a dictionary of possible compression sites for a nerve root,
        return a dictionary containing the relevant qualitative measurement for
        each level

        Output dictionary structure - { VLevel : measurement }

        POSS_CS - dictionary of possible nerve root compression sites
        SIDE - CS.SIDE object indicating nerve root side
        """

        qual_values = dict()
        ## Iterate through possible compression sites and add the relevant
        ## measurement for each site to the dictionary.
        for level, cs in poss_cs.items():
            if cs == CS.NONE:
                qual_values[level] = np.nan
            elif cs == CS.CENTRAL:
                qual_values[level] = self.schizas[level]
            elif cs == CS.LATERAL:
                qual_values[level] = self.lgrade[side][level]
            elif cs == CS.FORAMINAL:
                qual_values[level] = self.fgrade[side][level]
            else:
                raise RuntimeError("Invalid CS value.")
        return(qual_values)


class Participant():
    """Stores relevant data for each Participant/MRI. Provides functions to
    assess the number of nerve roots compressed for a given threshold."""

    def __init__(self, id_num, data, combn_rule = CombnRule.AND):
        """Creates a new participant, identified by an blinding ID (ID_NUM), and
        associated with an MRI scan whose measurments are contained within a row
        of the DATA frame.

        ID_NUM - blinding number of MRI for participant
        DATA - data frame containing MRI scans in wide format
        COMBN_RULE - CombnRule object"""

        self.id_num = id_num                                # Set blindingID
        self.measurements = Measurements(id_num, data)     # Create measurement set
        self.update_combn_rule(combn_rule)                  # Set combn_rule

        ## Generate simulated nerve roots for participant
        nerve_roots = list()
        for nlevel in NLevel:
            nerve_roots.append(NerveRoot(Side.LEFT, nlevel, self.measurements))
            nerve_roots.append(NerveRoot(Side.RIGHT, nlevel, self.measurements))
        self.nerve_roots = nerve_roots

    def debug_print(self):
        """Print all debug information for participant"""

        print(f"ID: {self.id_num}\n")
        self.measurements.debug_print()
        print("")
        for root in self.nerve_roots:
            root.debug_print()

    def check_compression(self, threshold, debug = False, csv_output = False):
        """Returns number of single and multiply compressed nerve roots for a
        given threshold. All nerve roots considered multiply compressed are also
```

```
        counted in single compressed nerve roots.

        THRESHOLD - Threshold dictionary specifying thresholds to consider a
        nerve root compressed.
        DEBUG - If true prints debug information for each assessed nerve root.
        """

        ## Set counters to zero
        self.num_roots_single_compres = 0
        self.num_roots_multi_compres = 0
        ## Iterate through nerve roots updating counters for each
        for root in self.nerve_roots:
            num_sites_compressed = \
              root.check_compression(threshold, self.combn_rule, debug)
            if num_sites_compressed >= 1:                # Compressed at least once
                self.num_roots_single_compres += 1
            if num_sites_compressed > 1:                 # Compressed more than once
                self.num_roots_multi_compres += 1

        if not csv_output:
            output = (self.num_roots_single_compres,
                      self.num_roots_multi_compres)
        else:
            output = self._csv_output(threshold,
                                      self.num_roots_single_compres,
                                      self.num_roots_multi_compres)
        return(output)

    def update_combn_rule(self, combn_rule):
        """Changes combination rule associated with participant. If an invalid
        combination rule is provided returns warning and does not change
        rule."""

        if combn_rule in CombnRule:
            self.combn_rule = combn_rule
        else:
            raise RuntimeWarning("Invalid combn_rule, ignoring.")

    def _csv_output(self, threshold, single, multi):
        """Create output list for return from the check_compression function."""

        output = ""
        ## Iterate through all threshold values.
        for value in threshold.values():
            output = output + f",{value}"
        ## Add single and multiple values
        output = output + f",{single},{multi}"
        ## Return output
        return(output)


class Threshold():
    """Iterator that provides all potential combination of values (for set step
    sizes) for provided ranges of measurements of the spinal canal. On
    completion returns None."""

    def __init__(self, dscsa_range, lrd_range, nfd_range,
                 schizas_range, lgrade_range, fgrade_range,
                 combn_rule):
        """Creates a new Threshold iterator.

        Expects range-objects as inputs for each variable."""

        ## Store ranges
        if not combn_rule == CombnRule.QUAL:
            self.dscsa_range = dscsa_range
            self.lrd_range = lrd_range
            self.nfd_range = nfd_range
        else:
            self.dscsa_range = range(0, 1, 1)
            self.lrd_range = range(0, 1, 1)
            self.nfd_range = range(0, 1, 1)
        if not combn_rule == CombnRule.QUANT:
            self.schizas_range = schizas_range
            self.lgrade_range = lgrade_range
            self.fgrade_range = fgrade_range
        else:
            self.schizas_range = range(0, 1, 1)
```

```
            self.lgrade_range = range(0, 1, 1)
            self.fgrade_range = range(0, 1, 1)

        ## Set up index variables storing the reached position within each
        ## variable range
        self.dscsa_index = -1
        self.lrd_index = 0
        self.nfd_index = 0
        self.schizas_index = 0
        self.lgrade_index = 0
        self.fgrade_index = 0

        ## Calculate total number of thresholds in iterator
        self.total = len(self.dscsa_range) * len(self.lrd_range) * \
                     len(self.nfd_range) * len(self.schizas_range) * \
                     len(self.lgrade_range) * len(self.fgrade_range)

    def __iter__(self):
        """Return the iterator."""
        return self

    def __next__(self):
        """Provide the next threshold as a dictionary of threshold values"""

        # Increment base index
        self.dscsa_index += 1

        # Detect overflow of each index and adjust the remainder of the indices
        # accordingly
        if (self.dscsa_index >= len(self.dscsa_range)):
            self.dscsa_index = 0
            self.lrd_index += 1
        if (self.lrd_index >= len(self.lrd_range)):
            self.lrd_index = 0
            self.nfd_index += 1
        if (self.nfd_index >= len(self.nfd_range)):
            self.nfd_index = 0
            self.schizas_index += 1
        if (self.schizas_index >= len(self.schizas_range)):
            self.schizas_index = 0
            self.lgrade_index += 1
        if (self.lgrade_index >= len(self.lgrade_range)):
            self.lgrade_index = 0
            self.fgrade_index += 1
        if (self.fgrade_index >= len(self.fgrade_range)):
            raise StopIteration   # Complete

        return({"dscsa" : self.dscsa_range[self.dscsa_index],
                "lrd" : self.lrd_range[self.lrd_index],
                "nfd" : self.nfd_range[self.nfd_index],
                "schizas" : self.schizas_range[self.schizas_index],
                "lgrade" : self.lgrade_range[self.lgrade_index],
                "fgrade" : self.fgrade_range[self.fgrade_index]})


def main(data, output_path, skip = []):
    """Main function."""

    ## Print current time
    current_time = datetime.now().strftime("%H:%M:%S")
    print(f"Starting at {current_time}")

    ## Create a participant object for each patient in the MRI data
    participants = list()
    for id_num in data.blindingID:
        participants.append(Participant(id_num, data))
    print("Completed participant generation")

    run = -1     # Keeps track of total number of iterations

    ## Iterate through all combination rules
    for combn_rule in CombnRule:

        ## Check if combination rule should be skipped
        if combn_rule in skip:
            print(f"Skipping combination {combn_rule.name}")
            continue
```

```
        ## Set combination rule across all participants
        print(f"Updating participants to use combination rule {combn_rule.name}")
        for participant in participants:
            participant.update_combn_rule(combn_rule)

        ## Create a new set of thresholds for the combination rule
        thresholds = Threshold(dscsa_range = DSCSA_RANGE,
                               lrd_range = LRD_RANGE,
                               nfd_range = NFD_RANGE,
                               schizas_range = SCHIZAS_RANGE,
                               lgrade_range = LGRADE_RANGE,
                               fgrade_range = FGRADE_RANGE,
                               combn_rule = combn_rule)

        ## Print total number of iterations expected for this threshold set
        print(f"Total expected runs: {thresholds.total}")

        ## Open a file to store the data in CSV format
        print(f"Creating {combn_rule.name}.csv")
        output_file = output_path + f"/{combn_rule.name}.csv"
        with open(output_file, 'w') as f:
            ## Column names header
            col_names = "run,blindingID,dscsa,lrd,nfd,schizas,lgrade,fgrade,sing,mult\n"
            f.write(col_names)

            ## Iterate through all threshold values
            for threshold in progressbar(thresholds, max_value=thresholds.total):
                run += 1       # Update run counter
                ## Iterate through all participants, checking compression status
                for participant in participants:
                    output = participant.check_compression(threshold, csv_output=True)
                    output = f"{run},{participant.id_num}" + output + "\n"
                    f.write(output)

    ## Print completion time
    current_time = datetime.now().strftime("%H:%M:%S")
    print(f"Completed at {current_time}")


## Import data
os.chdir(os.path.expanduser("~/Projects/phd_thesis/Analysis"))
DATA = pd.read_csv("wide_data/wide_data_v4.csv")
OUTPUT_PATH = "custom_algorithm"

DSCSA_RANGE = np.arange(40, 130, 10)
LRD_RANGE = np.arange(0, 7, 1)
NFD_RANGE = np.arange(4, 9, 1)
SCHIZAS_RANGE = np.arange(0, 4, 1)
LGRADE_RANGE = np.arange(0, 4, 1)
FGRADE_RANGE = np.arange(0, 4, 1)

if __name__ == "__main__":
    main(DATA, OUTPUT_PATH, [])
```

Bibliography

Abbas, J. et al. (2017). 'In the Quest for Degenerative Lumbar Spinal Stenosis Etiology: The Schmorl's Nodes Model'. In: *BMC Musculoskeletal Disorders* 18.1.

Adams, M. A. and Hutton, W. C. (Apr. 1983). 'The Mechanical Function of the Lumbar Apophyseal Joints'. In: *Spine* 8.3, pp. 327–330. pmid: 6623200.

Ahn, T.-J. et al. (2009). 'Effect of Intervertebral Disk Degeneration on Spinal Stenosis During Magnetic Resonance Imaging With Axial Loading'. In: *Neurologia medico-chirurgica* 49.6, pp. 242–247.

Albert, H. B. et al. (Apr. 2013). 'Antibiotic Treatment in Patients with Chronic Low Back Pain and Vertebral Bone Edema (Modic Type 1 Changes): A Double-Blind Randomized Clinical Controlled Trial of Efficacy'. In: *European Spine Journal* 22.4, pp. 697–707.

Alicioglu, B. et al. (Dec. 2012). 'Magnetic Resonance Imaging Predictors of Surgical Outcome in Degenerative Lumbar Spinal Stenosis'. In: *Japanese Journal of Radiology* 30.10, pp. 811–818.

Allaire, J. and Chollet, F. (2020a). *keras: R Interface to 'Keras'*. R package version 2.3.0.0.

Allaire, J. and Tang, Y. (2020b). *tensorflow: R Interface to 'TensorFlow'*. R package version 2.2.0.

Alsaleh, K. et al. (Feb. 2017). 'Radiographic Assessment of Degenerative Lumbar Spinal Stenosis: Is MRI Superior to CT?' In: *European Spine Journal* 26.2, pp. 362–367.

Aly, T. and Amin, O. (1st Feb. 2013). 'Geometrical Dimensions and Morphological Study of the Lumbar Spinal Canal in the Normal Egyptian Population'. In: *Orthopedics* 36.2, e229–e234.

Ammendolia, C. et al. (30th Aug. 2013). 'Nonoperative Treatment for Lumbar Spinal Stenosis with Neurogenic Claudication'. In: *Cochrane Database of Systematic Reviews*. Ed. by Cochrane Back and Neck Group.

Amonoo-Kuofi, H. S. (Jan. 1985). 'The Sagittal Diameter of the Lumbar Vertebral Canal in Normal Adult Nigerians'. In: *Journal of Anatomy* 140 (Pt 1), pp. 69–78. pmid: 4066472.

Amundsen, T. et al. (15th May 1995). 'Lumbar Spinal Stenosis. Clinical and Radiologic Features'. In: *Spine* 20.10, pp. 1178–1186. pmid: 7638662.

Andreisek, G., Hodler, J. and Steurer, J. (Dec. 2011). 'Uncertainties in the Diagnosis of Lumbar Spinal Stenosis'. In: *Radiology* 261.3, pp. 681–684.

Andreisek, G. et al. (Nov. 2013). 'A Systematic Review of Semiquantitative and Qualitative Radiologic Criteria for the Diagnosis of Lumbar Spinal Stenosis'. In: *American Journal of Roentgenology* 201.5, W735–W746.

Andreisek, G. et al. (Dec. 2014). 'Consensus Conference on Core Radiological Parameters to Describe Lumbar Stenosis - an Initiative for Structured Reporting'. In: *European Radiology* 24.12, pp. 3224–3232.

Arnold, J. B. (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.2.0.

Arnoldi, C. C. et al. (Mar. 1976). 'Lumbar Spinal Stenosis and Nerve Root Entrapment Syndromes:' in: *Clinical Orthopaedics and Related Research* &NA;115, 4???5.

Atilla, B. et al. (1st Nov. 1997). 'The Shape of the Lumbar Vertebral Canal in Newborns'. In: *Spine* 22.21, pp. 2469–2472. pmid: 9383851.

Azimi, P., Mohammadi, H. R. and Montazeri, A. (Dec. 2012). 'An Outcome Measure of Functionality in Patients with Lumber Spinal Stenosis: A Validation Study of the Iranian Version of Neurogenic Claudication Outcome Score (NCOS)'. In: *BMC Neurology* 12.1, p. 101.

Babyak, M. A. (1st May 2004). 'What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models'. In: *Psychosomatic Medicine* 66.3, pp. 411–421.

Bagley, C. et al. (31st Jan. 2019). 'Current Concepts and Recent Advances in Understanding and Managing Lumbar Spine Stenosis'. In: *F1000Research* 8, p. 137.

Barbaix, E. et al. (Mar. 1996). 'Anterior Sacrodural Attachments–Trolard's Ligaments Revisited'. In: *Manual Therapy* 1.2, pp. 88–91.

Bartko, J. J. (1976). 'On Various Intraclass Correlation Reliability Coefficients'. In: *Psychological Bulletin* 83.5, pp. 762–765.

Bartynski, W. S. and Lin, L. (2003). 'Lumbar Root Compression in the Lateral Recess: MR Imaging, Conventional Myelography, and CT Myelography Comparison with Surgical Confirmation'. In: *AJNR* 24, pp. 348–360.

Barz, T. et al. (May 2008). 'The Diagnostic Value of a Treadmill Test in Predicting Lumbar Spinal Stenosis'. In: *European Spine Journal* 17.5, pp. 686–690.

Barz, T. et al. (Apr. 2010). 'Nerve Root Sedimentation Sign: Evaluation of a New Radiological Sign in Lumbar Spinal Stenosis'. In: *Spine* 35.8, pp. 892–897.

Barz, T. et al. (Apr. 2014). 'Clinical Validity of the Nerve Root Sedimentation Sign in Patients with Suspected Lumbar Spinal Stenosis'. In: *The Spine Journal* 14.4, pp. 667–674.

Battié, M. C. et al. (Dec. 2014). 'Brief Report: Lumbar Spinal Stenosis Is a Highly Genetic Condition Partly Mediated by Disc Degeneration: Lumbar Spinal Stenosis Is Highly Genetic'. In: *Arthritis & Rheumatology* 66.12, pp. 3505–3510.

Beattie, P. F. et al. (Apr. 2000). 'Associations Between Patient Report of Symptoms and Anatomic Impairment Visible on Lumbar Magnetic Resonance Imaging:' in: *Spine* 25.7, pp. 819–828.

Beck, A. T. (1st June 1961). 'An Inventory for Measuring Depression'. In: *Archives of General Psychiatry* 4.6, p. 561.

Beers, G. et al. (Feb. 1985). 'Interobserver Discrepancies in Distance Measurements from Lumbar Spine CT Scans'. In: *American Journal of Roentgenology* 144.2, pp. 395–398.

Bennell, K., Dobson, F. and Hinman, R. (Nov. 2011). 'Measures of Physical Performance Assessments: Self-Paced Walk Test (SPWT), Stair Climb Test (SCT), Six-Minute Walk Test (6MWT), Chair Stand Test (CST), Timed Up & Go (TUG), Sock Test, Lift and Carry Test (LCT), and Car Task'. In: *Arthritis Care & Research* 63 Suppl 11, S350–370. pmid: 22588756.

Beresford, Z. M., Kendall, R. W. and Willick, S. E. (Jan. 2010). 'Lumbar Facet Syndromes:' in: *Current Sports Medicine Reports* 9.1, pp. 50–56.

Bibby, S. R. et al. (Dec. 2001). 'The Pathophysiology of the Intervertebral Disc'. In: *Joint Bone Spine* 68.6, pp. 537–542.

Boden, S. D. et al. (Mar. 1990). 'Abnormal Magnetic-Resonance Scans of the Lumbar Spine in Asymptomatic Subjects. A Prospective Investigation'. In: *The Journal of Bone and Joint Surgery. American Volume* 72.3, pp. 403–408. pmid: 2312537.

Bogduk, N. (2012). *Clinical and Radiological Anatomy of the Lumbar Spine*. 5th Edition. Churchill Livingstone.

Bolender, N. F., Schönström, N. S. and Spengler, D. M. (Feb. 1985). 'Role of Computed Tomography and Myelography in the Diagnosis of Central Spinal Stenosis.:' in: *The Journal of Bone & Joint Surgery* 67.2, pp. 240–246.

Boos, N. et al. (15th Dec. 1995). 'The Diagnostic Accuracy of Magnetic Resonance Imaging, Work Perception, and Psychoso-

cial Factors in Identifying Symptomatic Disc Herniations'. In: *Spine* 20.24, pp. 2613–2625. pmid: 8747239.
Bose, K. and Balasubramaniam, P. (1984). 'Nerve Root Canals of the Lumbar Spine'. In: *Spine* 9.1, pp. 16–18. pmid: 6719252.
Bowen, G. et al. (Dec. 2017). 'Leonardo Da Vinci (1452–1519) and His Depictions of the Human Spine'. In: *Child's Nervous System* 33.12, pp. 2067–2070.
Bramwell, B. (1886). *The Diseases of the Spinal Cord*. 2nd ed. New York. William Wood & Co. 201-254.
Bråten, L. C. H. et al. (16th Oct. 2019). 'Efficacy of Antibiotic Treatment in Patients with Chronic Low Back Pain and Modic Changes (the AIM Study): Double Blind, Randomised, Placebo Controlled, Multicentre Trial'. In: *BMJ*, p. l5654.
Breiman, L., ed. (1998). *Classification and Regression Trees*. Repr. Boca Raton: Chapman & Hall [u.a.] 358 pp.
Brinjikji, W. et al. (Dec. 2015a). 'MRI Findings of Disc Degeneration Are More Prevalent in Adults with Low Back Pain than in Asymptomatic Controls: A Systematic Review and Meta-Analysis'. In: *American Journal of Neuroradiology* 36.12, pp. 2394–2399.
Brinjikji, W. et al. (Apr. 2015b). 'Systematic Literature Review of Imaging Features of Spinal Degeneration in Asymptomatic Populations'. In: *American Journal of Neuroradiology* 36.4, pp. 811–816.
Brown, H. A. (1938). 'Enlargement of the Ligamentum Flavum'. In: *The Journal of Bone & Joint Surgery* 20.2, pp. 325–338.
Burgstaller, J. M. et al. (Sept. 2016). 'Is There an Association Between Pain and Magnetic Resonance Imaging Parameters in Patients With Lumbar Spinal Stenosis?:' in: *SPINE* 41.17, E1053–E1062.
Carleton, R. N. et al. (1st Mar. 2013). 'The Center for Epidemiologic Studies Depression Scale: A Review with a Theoretical and Empirical Examination of Item Content and Factor Structure'. In: *PLoS ONE* 8.3. Ed. by H. R. Baradaran, e58067.
Cavanaugh, J. M. (15th Aug. 1995). 'Neural Mechanisms of Lumbar Pain'. In: *Spine* 20.16, pp. 1804–1809. pmid: 7502138.
Chaput, C. et al. (Aug. 2007). 'The Significance of Increased Fluid Signal on Magnetic Resonance Imaging in Lumbar Facets in Relationship to Degenerative Spondylolisthesis:' in: *Spine* 32.17, pp. 1883–1887.
Charcot, J. (1859). '2nd Series, 5, Part 2'. In: *Comptes Rendus Des Seances et Memoires de La Societe de Biologie*, pp. 225–238.
Chatha, D. S. and Schweitzer, M. E. (2011). 'MRI Criteria of Developmental Lumbar Spinal Stenosis Revisited'. In: *Bulletin of the NYU hospital for joint diseases* 69.4, pp. 303–307. pmid: 22196386.
Chen, C. et al. (15th May 1997). 'Effects of Phospholipase A2 on Lumbar Nerve Root Structure and Function'. In: *Spine* 22.10, pp. 1057–1064. pmid: 9160462.
Cheung, J. P.-Y., Shigematsu, H. and Cheung, K. M.-C. (Aug. 2014a). 'Verification of Measurements of Lumbar Spinal Dimensions in T1- and T2-Weighted Magnetic Resonance Imaging Sequences'. In: *The Spine Journal* 14.8, pp. 1476–1483.
Cheung, J. P.-Y. et al. (June 2014b). 'Defining Clinically Relevant Values for Developmental Spinal Stenosis: A Large-Scale Magnetic Resonance Imaging Study'. In: *Spine* 39.13, pp. 1067–1076.
Cheung, K. M. C. et al. (Apr. 2009). 'Prevalence and Pattern of Lumbar Magnetic Resonance Imaging Changes in a Population Study of One Thousand Forty-Three Individuals:' in: *Spine* 34.9, pp. 934–940.
Chun, S.-W. et al. (Jan. 2017). 'Cerebrospinal Fluid Dynamics at the Lumbosacral Level in Patients with Spinal Stenosis: A Pilot Study: CSF DYNAMICS IN SPINAL STENOSIS'. In: *Journal of Orthopaedic Research* 35.1, pp. 104–112.
Chung, S. S. et al. (20th Apr. 2000). 'Effect of Low Back Posture on the Morphology of the Spinal Canal'. In: *Skeletal Radiology* 29.4, pp. 217–223.
Ciol, M. A. et al. (Mar. 1996). 'An Assessment of Surgery for Spinal Stenosis: Time Trends, Geographic Variations, Complications, and Reoperations'. In: *Journal of the American Geriatrics Society* 44.3, pp. 285–290.
Ciric, I. et al. (Oct. 1980). 'The Lateral Recess Syndrome. A Variant of Spinal Stenosis'. In: *Journal of Neurosurgery* 53.4, pp. 433–443. pmid: 7420163.
Clarençon, F. et al. (2016). 'The Degenerative Spine'. In: *Magnetic Resonance Imaging Clinics of North America* 24.3, pp. 495–513.
Clark, G. A., Panjabi, M. M. and Wetzel, F. T. (Mar. 1985). 'Can Infant Malnutrition Cause Adult Vertebral Stenosis?' In: *Spine* 10.2, pp. 165–170. pmid: 4002040.
Cohen, J. (Apr. 1960). 'A Coefficient of Agreement for Nominal Scales'. In: *Educational and Psychological Measurement* 20.1, pp. 37–46.
Comer, C. M. et al. (Dec. 2009). 'Assessment and Management of Neurogenic Claudication Associated with Lumbar Spinal Stenosis in a UK Primary Care Musculoskeletal Service: A Survey of Current Practice among Physiotherapists'. In: *BMC Musculoskeletal Disorders* 10.1, p. 121.
Conway, J., Tomkins, C. C. and Haig, A. J. (Sept. 2011). 'Walking Assessment in People with Lumbar Spinal Stenosis: Capacity, Performance, and Self-Report Measures'. In: *The Spine Journal* 11.9, pp. 816–823.
Cowley, P. (2016). 'Neuroimaging of Spinal Canal Stenosis'. In: *Magnetic Resonance Imaging Clinics of North America* 24.3, pp. 523–539.
Cramer, G. D. et al. (Mar. 2003). 'Dimensions of the Lumbar Intervertebral Foramina as Determined from the Sagittal Plane Magnetic Resonance Imaging Scans of 95 Normal Subjects'. In: *Journal of Manipulative and Physiological Therapeutics* 26.3, pp. 160–170.
Cummins, J. et al. (1st Apr. 2006). 'Descriptive Epidemiology and Prior Healthcare Utilization of Patients in the Spine Patient Outcomes Research Trial's (SPORT) Three Observational Cohorts: Disc Herniation, Spinal Stenosis, and Degenerative Spondylolisthesis'. In: *Spine* 31.7, pp. 806–814. pmid: 16582855.
Cybenko, G. (Dec. 1989). 'Approximation by Superpositions of a Sigmoidal Function'. In: *Mathematics of Control, Signals, and Systems* 2.4, pp. 303–314.
Danielson, B. I. et al. (Nov. 1998). 'Axial Loading of the Spine during CT and MR in Patients with Suspected Lumbar Spinal Stenosis'. In: *Acta Radiologica* 39.6, pp. 604–611.
Danielson, B. and Willén, J. (Dec. 2001). 'Axially Loaded Magnetic Resonance Image of the Lumbar Spine in Asymptomatic Individuals:' in: *Spine* 26.23, pp. 2601–2606.
Deen, H. G. et al. (Oct. 2000). 'Test-Retest Reproducibility of the Exercise Treadmill Examination in Lumbar Spinal Stenosis'. In: *Mayo Clinic Proceedings* 75.10, pp. 1002–1007. pmid: 11040847.
Dejerine, J. (1911). *Semiologie Des Affections Du Systeme Nerveux*. Paris, Masson et Cie. 267 pp.
De Roos, A. et al. (Sept. 1987). 'MR Imaging of Marrow Changes Adjacent to End Plates in Degenerative Lumbar Disk Disease'. In: *American Journal of Roentgenology* 149.3, pp. 531–534.
De Schepper, E. I. T. et al. (Apr. 2013). 'Diagnosis of Lumbar Spinal Stenosis: An Updated Systematic Review of the Accuracy of Diagnostic Tests'. In: *Spine* 38.8, E469–E481.
De Winter, J. C. F., Gosling, S. D. and Potter, J. (2016). 'Comparing the Pearson and Spearman Correlation Coefficients across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data.' In: *Psychological Methods* 21.3, pp. 273–290.
Deyo, R. A. et al. (15th Sept. 1998). 'Outcome Measures for Low Back Pain Research. A Proposal for Standardized Use'. In: *Spine* 23.18, pp. 2003–2013. pmid: 9779535.
Deyo, R. A. et al. (2010). 'Trends, Major Medical Complications, and Charges Associated With Surgery for Lumbar Spinal Stenosis in Older Adults'. In: p. 7.
Dincer, F. et al. (1991). 'Lateral Recess Syndrome and Computed Tomography'. In: *Turkish Neurosurgery* 2.1.
Downie, W. W. et al. (1st Aug. 1978). 'Studies with Pain Rating Scales.' In: *Annals of the Rheumatic Diseases* 37.4, pp. 378–381.
Doyle, A. J. and Merrilees, M. (Apr. 2004). 'Synovial Cysts of the Lumbar Facet Joints in a Symptomatic Population: Prevalence on Magnetic Resonance Imaging'. In: *Spine* 29.8, pp. 874–878.
Dreyer, S. J. and Dreyfuss, P. H. (Mar. 1996). 'Low Back Pain and the Zygapophysial (Facet) Joints'. In: *Archives of Physical Medicine and Rehabilitation* 77.3, pp. 290–300.
Dwyer, T., Couper, D. and Walter, S. D. (May 2001). 'Sources of Heterogeneity in the Meta-Analysis of Observational Studies The Example of SIDS and Sleeping Position'. In: *Journal of Clinical Epidemiology* 54.5, pp. 440–447.
Eisenstein, S. (May 1977). 'The Morphometry and Pathological Anatomy of the Lumbar Spine in South African Negroes and Caucasoids with Specific Reference to Spinal Stenosis'. In: *The*

Journal of Bone and Joint Surgery. British volume 59-B.2, pp. 173–180.
Elsberg, C. A. (1913). 'Experiences in Spinal Surgery'. In: *Surg. Gynae. & Obst.* 16, pp. 117–132.
Eun, S. S. et al. (May 2012). 'MRI versus CT for the Diagnosis of Lumbar Spinal Stenosis'. In: *Journal of Neuroradiology* 39.2, pp. 104–109.
EuroQol Group (Dec. 1990). 'EuroQol–a New Facility for the Measurement of Health-Related Quality of Life'. In: *Health Policy (Amsterdam, Netherlands)* 16.3, pp. 199–208. pmid: 10109801.
Everitt, B. and Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. 4th ed. Cambridge, UK ; New York: Cambridge University Press. 468 pp.
Fairbank, J. and Pynsent, P. (2000). 'The Oswestry Disability Index'. In: 25.22, pp. 2940–2953.
Fang, D. et al. (Aug. 1994). 'Computed Tomographic Osteometry of the Asian Lumbar Spine'. In: *Journal of Spinal Disorders* 7.4, pp. 307–316. pmid: 7949698.
Fanuele, J. C. et al. (June 2000). 'The Impact of Spinal Problems on the Health Status of Patients: Have We Underestimated the Effect?' In: *Spine* 25.12, pp. 1509–1514.
Fardon, D. F. et al. (Nov. 2014). 'Lumbar Disc Nomenclature: Version 2.0'. In: *The Spine Journal* 14.11, pp. 2525–2545.
Field, A. P., Miles, J. and Field, Z. (2012). *Discovering Statistics Using R*. London ; Thousand Oaks, Calif: Sage. 957 pp.
Fisher, R. A. (Sept. 1936). 'THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS'. In: *Annals of Eugenics* 7.2, pp. 179–188.
Fluss, R., Faraggi, D. and Reiser, B. (Aug. 2005). 'Estimation of the Youden Index and Its Associated Cutoff Point'. In: *Biometrical Journal* 47.4, pp. 458–472.
Fraser, S., Roberts, L. and Murphy, E. (Nov. 2009). 'Cauda Equina Syndrome: A Literature Review of Its Definition and Clinical Presentation'. In: *Archives of Physical Medicine and Rehabilitation* 90.11, pp. 1964–1968.
Friedman, J., Hastie, T. and Tibshirani, R. (2010). 'Regularization Paths for Generalized Linear Models via Coordinate Descent'. In: *Journal of statistical software* 33.1, p. 1.
Fujiwara, A. et al. (26th Oct. 1999). 'The Relationship between Facet Joint Osteoarthritis and Disc Degeneration of the Lumbar Spine: An MRI Study'. In: *European Spine Journal* 8.5, pp. 396–401.
Fujiwara, A. et al. (July 2003). 'Association of the Japanese Orthopaedic Association Score With the Oswestry Disability Index, Roland-Morris Disability Questionnaire, and Short-Form 36:' in: *Spine* 28.14, pp. 1601–1607.
Fukusaki, M. et al. (June 1998). 'Symptoms of Spinal Stenosis Do Not Improve after Epidural Steroid Injection'. In: *The Clinical Journal of Pain* 14.2, pp. 148–151. pmid: 9647457.
Gagen, R. M., Parsons, C. and Hutchinson, C. (2017a). *MRI in degenerative lumbar spinal stenosis: is there a relationship between the dural-sac cross-sectional area and patient pain, disability and quality of life? A systematic review*. PROSPERO 2017 CRD42017064865 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42017064865.
— (2017b). *MRI in degenerative lumbar spinal stenosis: what MRI findings are associated with neurogenic claudication? A systematic review*. PROSPERO 2017 CRD42017064873 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42017064873.
Gagen, RM, Dragos, C and Hutchinson, CE (2020). 'Can MRI of the Lumbar Spine Identify Neurogenic Claudication Patients Reliably? - Comparison of Single Variable and Machine Learning Approaches.' In: *Proceedings of the International Society of Magnetic Resonance in Medicine*. ISMRM. Virtual.
Galbusera, F. et al. (2016). 'MR Imaging and Radiographic Imaging of Degenerative Spine Disorders and Spine Alignment'. In: *Magnetic Resonance Imaging Clinics of North America* 24.3, pp. 515–522.
Galen (2011a). 'De Nervorum Dissectione'. In: *Claudii Galeni Opera Omnia*. Ed. by K. G. Kühn. Vol. 2. Cambridge University Press, pp. 831–856.
— (2011b). 'De Ossibus Ad Tirones'. In: *Claudii Galeni Opera Omnia*. Ed. by K. G. Kühn. Vol. 2. Cambridge University Press, pp. 732–778.
Galen and May, M. T. (1968). *Galen on the Usefulness of the Parts of the Body*. Ithaca, N.Y: Cornell University Press.
Gardner, A., Gardner, E. and Morley, T. (May 2011). 'Cauda Equina Syndrome: A Review of the Current Clinical and Medico-Legal Position'. In: *European Spine Journal* 20.5, pp. 690–697.
Gardner, B. et al. (30th Aug. 2012). 'Towards Parsimony in Habit Measurement: Testing the Convergent and Predictive Validity of an Automaticity Subscale of the Self-Report Habit Index'. In: *The International Journal of Behavioral Nutrition and Physical Activity* 9, p. 102. pmid: 22935297.
Garfin, S. R., Rydevik, B. L. and Brown, R. A. (Feb. 1991). 'Compressive Neuropathy of Spinal Nerve Roots. A Mechanical or Biological Problem?' In: *Spine* 16.2, pp. 162–166. pmid: 1826376.
Geisser, M. E. et al. (Nov. 2007). 'Spinal Canal Size and Clinical Symptoms Among Persons Diagnosed With Lumbar Spinal Stenosis:' in: *The Clinical Journal of Pain* 23.9, pp. 780–785.
Ghasemi, A. et al. (Feb. 2016). 'The Relation Between Sacral Angle and Vertical Angle of Sacral Curvature and Lumbar Disc Degeneration: A Case–Control Study'. In: *Medicine* 95.6, e2746.
Gobbens, R. J. J. et al. (June 2010). 'The Tilburg Frailty Indicator: Psychometric Properties'. In: *Journal of the American Medical Directors Association* 11.5, pp. 344–355. pmid: 20511102.
Goni, V. G. et al. (2014). 'Comparison of the Oswestry Disability Index and Magnetic Resonance Imaging Findings in Lumbar Canal Stenosis: An Observational Study'. In: *Asian Spine Journal* 8.1, p. 44.
Gouzien, P. et al. (June 1990). 'Measurements of the Normal Lumbar Spinal Canal by Computed Tomography'. In: *Surgical and Radiologic Anatomy* 12.2, pp. 143–148.
Greenbarg, P. E. et al. (1988). 'Epidural Anesthesia for Lumbar Spine Surgery'. In: *Journal of Spinal Disorders* 1.2, pp. 139–143. pmid: 2980070.
Greenough, C. G. and Fraser, R. D. (Jan. 1992). 'Assessment of Outcome in Patients with Low-Back Pain'. In: *Spine* 17.1, pp. 36–41. pmid: 1531553.
Greenwell, B., Boehmke, B. and Gray, B. (2020). *vip: Variable Importance Plots*. R package version 0.2.2.
Grenier, N. et al. (Nov. 1987). 'Normal and Degenerative Posterior Spinal Structures: MR Imaging.' In: *Radiology* 165.2, pp. 517–525.
Grimes, P. F., Massie, J. B. and Garfin, S. R. (Aug. 2000). 'Anatomic and Biomechanical Analysis of the Lower Lumbar Foraminal Ligaments:' in: *Spine* 25.16, pp. 2009–2014.
Guralnik, J. M. et al. (Apr. 2000). 'Lower Extremity Function and Subsequent Disability: Consistency across Studies, Predictive Models, and Value of Gait Speed Alone Compared with the Short Physical Performance Battery'. In: *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 55.4, pp. M221–231. pmid: 10811152.
Guralnik, J. M. et al. (1994). 'A Short Physical Performance Battery Assessing Lower Extremity Function: Association with Self-Reported Disability and Prediction of Mortality and Nursing Home Admission'. In: *Journal of gerontology* 49.2, pp. M85–M94.
Haig, A. J. et al. (Dec. 2006a). 'Predictors of Pain and Function in Persons With Spinal Stenosis, Low Back Pain, and No Back Pain:' in: *Spine* 31.25, pp. 2950–2957.
Haig, A. J. et al. (July 2006b). 'Spinal Stenosis, Back Pain, or No Symptoms at All? A Masked Study Comparing Radiologic and Electrodiagnostic Diagnoses to the Clinical Impression'. In: *Archives of Physical Medicine and Rehabilitation* 87.7, pp. 897–903.
Haig, A. J. et al. (Feb. 2007). 'Electromyographic and Magnetic Resonance Imaging to Predict Lumbar Stenosis, Low-Back Pain, and No Back Symptoms'. In: *The Journal of Bone and Joint Surgery. American Volume* 89.2, pp. 358–366. pmid: 17272451.
Hamanishi, C. et al. (Oct. 1994). 'Cross-Sectional Area of the Stenotic Lumbar Dural Tube Measured from the Transverse Views of Magnetic Resonance Imaging'. In: *Journal of Spinal Disorders* 7.5, pp. 388–393. pmid: 7819638.
Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-Color Illustrations*. Springer Series in Statistics. New York: Springer. 533 pp.
Henderson, L. et al. (Aug. 2012). 'Is Spinal Stenosis Assessment Dependent on Slice Orientation? A Magnetic Resonance Imaging Study'. In: *European Spine Journal* 21.S6, pp. 760–764.

Herzog, R. J. et al. (June 1991). 'The Importance of Posterior Epidural Fat Pad in Lumbar Central Canal Stenosis'. In: *Spine* 16 (6 Suppl), S227–233. pmid: 1862417.

Hill, J. C. et al. (15th May 2008). 'A Primary Care Back Pain Screening Tool: Identifying Patient Subgroups for Initial Treatment'. In: *Arthritis & Rheumatism* 59.5, pp. 632–641.

Hinck, V. C., Clark, W. M. and Hopkins, C. E. (May 1966). 'Normal Interpediculate Distances (Minimum and Maximum) in Children and Adults'. In: *The American Journal of Roentgenology, Radium Therapy, and Nuclear Medicine* 97.1, pp. 141–153. pmid: 5938032.

Hinck, V. C., Hopkins, C. E. and Clark, W. M. (Nov. 1965a). 'Sagittal Diameter of the Lumbar Spinal Canal in Children and Adults'. In: *Radiology* 85.5, pp. 929–937.

— (Nov. 1965b). 'Sagittal Diameter of the Lumbar Spinal Canal in Children and Adults'. In: *Radiology* 85.5, pp. 929–937.

Hirasawa, Y. et al. (Feb. 2007). 'Postural Changes of the Dural Sac in the Lumbar Spines of Asymptomatic Individuals Using Positional Stand-Up Magnetic Resonance Imaging:' in: *Spine* 32.4, E136–E140.

Holte, R. C. (1993). 'Very Simple Classification Rules Perform Well on Most Commonly Used Datasets'. In: *Machine Learning*, p. 32.

Hong, J. H. et al. (2015). 'Does Spinal Stenosis Correlate with MRI Findings and Pain, Psychologic Factor and Quality of Life?' In: *Korean Journal of Anesthesiology* 68.5, p. 481.

Hou, Z.-j. et al. (2015). 'Changes in Lumbosacral Spinal Nerve Roots on Diffusion Tensor Imaging in Spinal Stenosis'. In: *Neural Regeneration Research* 10.11, p. 1860.

Houck, P. R. et al. (2002). 'Reliability of the Self-Report Version of the Panic Disorder Severity Scale'. In: *Depression and Anxiety* 15.4, pp. 183–185.

Hoy, D. et al. (June 2012). 'A Systematic Review of the Global Prevalence of Low Back Pain'. In: *Arthritis and Rheumatism* 64.6, pp. 2028–2037. pmid: 22231424.

Ikawa, M., Atsuta, Y. and Tsunekawa, H. (Nov. 2005). 'Ectopic Firing Due to Artificial Venous Stasis in Rat Lumbar Spinal Canal Stenosis Model: A Possible Pathogenesis of Neurogenic Intermittent Claudication'. In: *Spine* 30.21, pp. 2393–2397.

Inufusa, A. et al. (1st Nov. 1996). 'Anatomic Changes of the Spinal Canal and Intervertebral Foramen Associated with Flexion-Extension Movement'. In: *Spine* 21.21, pp. 2412–2420. pmid: 8923625.

Ishimoto, Y. et al. (Oct. 2012). 'Prevalence of Symptomatic Lumbar Spinal Stenosis and Its Association with Physical Performance in a Population-Based Cohort in Japan: The Wakayama Spine Study'. In: *Osteoarthritis and Cartilage* 20.10, pp. 1103–1108.

— (June 2013). 'Associations between Radiographic Lumbar Spinal Stenosis and Clinical Symptoms in the General Population: The Wakayama Spine Study'. In: *Osteoarthritis and Cartilage* 21.6, pp. 783–788.

Ishimoto, Y. et al. (June 2017). 'Association of Lumbar Spondylolisthesis With Low Back Pain and Symptomatic Lumbar Spinal Stenosis in a Population-Based Cohort: The Wakayama Spine Study'. In: *SPINE* 42.11, E666–E671.

Ito, T. et al. (15th Jan. 1996). 'Histologic Evidence of Absorption of Sequestration-Type Herniated Disc'. In: *Spine* 21.2, pp. 230–234. pmid: 8720409.

Izzo, R. et al. (Jan. 2013a). 'Biomechanics of the Spine. Part I: Spinal Stability'. In: *European Journal of Radiology* 82.1, pp. 118–126.

— (Jan. 2013b). 'Biomechanics of the Spine. Part II: Spinal Instability'. In: *European Journal of Radiology* 82.1, pp. 127–138.

James, G. et al., eds. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer. 426 pp.

Jensen, M. C. et al. (14th July 1994). 'Magnetic Resonance Imaging of the Lumbar Spine in People without Back Pain'. In: *New England Journal of Medicine* 331.2, pp. 69–73.

Jensen, Ole K (2019). 'Rapid Response: Efficacy of Antibiotic Treatment in Patients with Chronic Low Back Pain and Modic Changes (the AIM Study): Double Blind, Randomised, Placebo Controlled, Multicentre Trial'. In: *British Medical Journal* 361, p. l5654.

Jentzsch, T. et al. (2013). 'Lumbar Facet Joint Arthritis Is Associated with More Coronal Orientation of the Facet Joints at the Upper Lumbar Spine'. In: *Radiology Research and Practice* 2013, pp. 1–9.

Jiang, J. et al. (Mar. 2017). 'Multifidus Degeneration, A New Risk Factor for Lumbar Spinal Stenosis: A Case–Control Study'. In: *World Neurosurgery* 99, pp. 226–231.

Jinkins, J. R. (1993). 'MR of Enhancing Nerve Roots in the Unoperated Lumbosacral Spine'. In: *AJNR. American journal of neuroradiology* 14.1, pp. 193–202. pmid: 8427089.

Jönsson, B. et al. (15th Dec. 1997). 'A Prospective and Consecutive Study of Surgically Treated Lumbar Spinal Stenosis. Part I: Clinical Features Related to Radiographic Findings'. In: *Spine* 22.24, pp. 2932–2937. pmid: 9431629.

Kalff, R. et al. (13th Sept. 2013). 'Degenerative Lumbar Spinal Stenosis in Older People'. In: *Deutsches Aerzteblatt Online*.

Kalichman, L. and Hunter, D. J. (Oct. 2007). 'Lumbar Facet Joint Osteoarthritis: A Review'. In: *Seminars in Arthritis and Rheumatism* 37.2, pp. 69–80.

Kalichman, L. et al. (July 2009). 'Spinal Stenosis Prevalence and Association with Symptoms: The Framingham Study'. In: *The Spine Journal* 9.7, pp. 545–550.

Kamper, S. J., Maher, C. G. and Mackay, G. (2009). 'Global Rating of Change Scales: A Review of Strengths and Weaknesses and Considerations for Design'. In: *The Journal of Manual & Manipulative Therapy* 17.3, pp. 163–170. pmid: 20046623.

Kanno, H. et al. (Feb. 2012). 'Dynamic Change of Dural Sac Cross-Sectional Area in Axial Loaded Magnetic Resonance Imaging Correlates With the Severity of Clinical Symptoms in Patients With Lumbar Spinal Canal Stenosis:' in: *Spine* 37.3, pp. 207–213.

Kapural, L. et al. (Sept. 2007). 'Value of the Magnetic Resonance Imaging in Patients With Painful Lumbar Spinal Stenosis (LSS) Undergoing Lumbar Epidural Steroid Injections:' in: *The Clinical Journal of Pain* 23.7, pp. 571–575.

Karavelioglu, E. et al. (2016). 'Ligamentum Flavum Thickening at Lumbar Spine Is Associated with Facet Joint Degeneration: An MRI Study'. In: *Journal of Back and Musculoskeletal Rehabilitation* 29.4.

Katz, J. N. (2008). 'Lumbar Spinal Stenosis'. In: *The New England Journal of Medicine*, p. 8.

Katz, J. N. et al. (Sept. 1995). 'Degenerative Lumbar Spinal Stenosis Diagnostic Value of the History and Physical Examination'. In: *Arthritis & Rheumatism* 38.9, pp. 1236–1241.

Kaufman, S., Rosset, S. and Perlich, C. (2011). 'Leakage in Data Mining: Formulation, Detection, and Avoidance'. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*. The 17th ACM SIGKDD International Conference. San Diego, California, USA: ACM Press, p. 556.

Kauppila, L. I. et al. (1st Sept. 1998). 'Degenerative Displacement of Lumbar Vertebrae. A 25-Year Follow-up Study in Framingham'. In: *Spine* 23.17, 1868–1873, discussion 1873–1874. pmid: 9762744.

Kim, H.-J. et al. (2013). 'Gender Difference of Symptom Severity in Lumbar Spinal Stenosis: Role of Pain Sensitivity'. In: *Pain Physician*, p. 10.

Kim, H.-J. et al. (Feb. 2015). 'The Significance of Pain Catastrophizing in Clinical Manifestations of Patients with Lumbar Spinal Stenosis: Mediation Analysis with Bootstrapping'. In: *The Spine Journal* 15.2, pp. 238–246.

Kim, N. H. and Lee, J. W. (June 1995). 'The Relationship between Isthmic and Degenerative Spondylolisthesis and the Configuration of the Lamina and Facet Joints'. In: *European Spine Journal* 4.3, pp. 139–144.

Kim, Y. K. et al. (Oct. 2013). 'Diagnostic Advancement of Axial Loaded Lumbar Spine MRI in Patients With Clinically Suspected Central Spinal Canal Stenosis:' in: *Spine* 38.21, E1342–E1347.

Kim, Y. U. et al. (Oct. 2015). 'Clinical Symptoms of Lumbar Spinal Stenosis Associated with Morphological Parameters on Magnetic Resonance Images'. In: *European Spine Journal* 24.10, pp. 2236–2243.

Kim, Y. U. et al. (2017). 'The Role of the Ligamentum Flavum Area as a Morphological Parameter of Lumbar Central Spinal Stenosis'. In: *Pain Physician* 20, E4196–E424.

Kindler, E. and Krivy, I. (Apr. 2011). 'Object-Oriented Simulation of Systems with Sophisticated Control'. In: *International Journal of General Systems* 40.3, pp. 313–343.

Kitab, S. A., Alsulaiman, A. M. and Benzel, E. C. (May 2014). 'Anatomic Radiological Variations in Developmental Lumbar

Spinal Stenosis: A Prospective, Control-Matched Comparative Analysis'. In: *The Spine Journal* 14.5, pp. 808–815.
Knirsch, W. et al. (Apr. 2005). 'Normal Values of the Sagittal Diameter of the Lumbar Spine (Vertebral Body and Dural Sac) in Children Measured by MRI'. In: *Pediatric Radiology* 35.4, pp. 419–424.
Kobayashi, S. et al. (5th June 2008). 'Effects of Arterial Ischemia and Venous Congestion on the Lumbar Nerve Root in Dogs'. In: *Journal of Orthopaedic Research* 26.11, pp. 1533–1540.
Kobayashi, S. et al. (Oct. 2015). 'Circulatory Dynamics of the Cauda Equina in Lumbar Canal Stenosis Using Dynamic Contrast-Enhanced Magnetic Resonance Imaging'. In: *The Spine Journal* 15.10, pp. 2132–2141.
Koc, Z. et al. (May 2009). 'Effectiveness of Physical Therapy and Epidural Steroid Injections in Lumbar Spinal Stenosis:' in: *Spine* 34.10, pp. 985–989.
Kohlmann, T., Tabor, J. and Bovasso, E. (1994). 'Die patientennahe Diagnostik von Functionsein- schrankungen im Alltag.' In: *Psychomed* 6, pp. 21–27.
Kopec, J. A. et al. (1st Feb. 1995). 'The Quebec Back Pain Disability Scale. Measurement Properties'. In: *Spine* 20.3, pp. 341–352. pmid: 7732471.
Kornberg, M. and Rechtine, G. R. (May 1985). 'Quantitative Assessment of the Fifth Lumbar Spinal Canal by Computed Tomography in Symptomatic L4-L5 Disc Disease'. In: *Spine* 10.4, pp. 328–330. pmid: 4049093.
Kreiner, D. S. et al. (July 2013). 'An Evidence-Based Clinical Guideline for the Diagnosis and Treatment of Degenerative Lumbar Spinal Stenosis (Update)'. In: *The Spine Journal* 13.7, pp. 734–743.
Kubosch, D. et al. (Aug. 2015). 'The Lumbar Spine as a Dynamic Structure Depicted in Upright MRI:' in: *Medicine* 94.32, e1299.
Kuhn, M. (2008). 'Building Predictive Models in R Using the Caret Package'. In: *Journal of Statistical Software* 28.5.
— (2020). *caret: Classification and Regression Training*. R package version 6.0-86.
Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer. 600 pp.
Kuhn, M. and Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*
Kuittinen, P. et al. (Dec. 2012). 'Accuracy and Reproducibility of a Retrospective Outcome Assessment for Lumbar Spinal Stenosis Surgery'. In: *BMC Musculoskeletal Disorders* 13.1, p. 83.
Kuittinen, P. et al. (2014a). 'Correlation of Lateral Stenosis in MRI with Symptoms, Walking Capacity and EMG Findings in Patients with Surgically Confirmed Lateral Lumbar Spinal Canal Stenosis'. In: p. 6.
Kuittinen, P. et al. (17th Sept. 2014b). 'Preoperative MRI Findings Predict Two-Year Postoperative Clinical Outcome in Lumbar Spinal Stenosis'. In: *PLoS ONE* 9.9. Ed. by P. Arnold, e106404.
Kuittinen, P. et al. (Dec. 2014c). 'Visually Assessed Severity of Lumbar Spinal Canal Stenosis Is Paradoxically Associated with Leg Pain and Objective Walking Ability'. In: *BMC Musculoskeletal Disorders* 15.1, p. 348.
Kuorinka, I. et al. (1987). 'Standardised Nordic Questionnaires for the Analysis of Musculoskeletal Symptoms'. In: *Applied ergonomics* 18.3, pp. 233–237.
Kuslich, S. D., Ulstrom, C. L. and Michael, C. J. (Apr. 1991). 'The Tissue Origin of Low Back Pain and Sciatica: A Report of Pain Response to Tissue Stimulation during Operations on the Lumbar Spine Using Local Anesthesia'. In: *The Orthopedic Clinics of North America* 22.2, pp. 181–187. pmid: 1826546.
Laidlaw, K. et al. (Apr. 2007). 'The Attitudes to Ageing Questionnaire (AAQ): Development and Psychometric Properties'. In: *International Journal of Geriatric Psychiatry* 22.4, pp. 367–379.
Lamb, S. E. et al. (Sept. 2005). 'Development of a Common Outcome Data Set for Fall Injury Prevention Trials: The Prevention of Falls Network Europe Consensus'. In: *Journal of the American Geriatrics Society* 53.9, pp. 1618–1622. pmid: 16137297.
Lau, Y. Y. O. et al. (Oct. 2017). 'Changes in Dural Sac Caliber with Standing MRI Improve Correlation with Symptoms of Lumbar Spinal Stenosis'. In: *European Spine Journal* 26.10, pp. 2666–2675.
Laudato, P. A., Kulik, G. and Schizas, C. (Oct. 2015). 'Relationship between Sedimentation Sign and Morphological Grade in Symptomatic Lumbar Spinal Stenosis'. In: *European Spine Journal* 24.10, pp. 2264–2268.
Laurencin, C. T. et al. (1999). 'The Stenosis Ratio: A New Tool for the Diagnosis of Degenerative Spinal Stenosis'. In: *International Journal of Surgical Investigation* 1.2, pp. 127–131. pmid: 11341632.
Lee, B. C., Kazam, E. and Newman, A. D. (July 1978). 'Computed Tomography of the Spine and Spinal Cord'. In: *Radiology* 128.1, pp. 95–102.
Lee, C. K., Hansen, H. T. and Weiss, A. B. (Sept. 1978). 'Developmental Lumbar Spinal Stenosis. Pathology and Surgical Treatment'. In: *Spine* 3.3, pp. 246–255. pmid: 715553.
Lee, C. K., Rauschning, W. and Glenn, W. (Mar. 1988). 'Lateral Lumbar Spinal Canal Stenosis: Classification, Pathologic Anatomy and Surgical Decompression'. In: *Spine* 13.3, pp. 313–320. pmid: 3388117.
Lee, H. M. et al. (1st Aug. 1995). 'Morphometric Study of the Lumbar Spinal Canal in the Korean Population'. In: *Spine* 20.15, pp. 1679–1684. pmid: 7482017.
Lee, S. et al. (Apr. 2010). 'A Practical MRI Grading System for Lumbar Foraminal Stenosis'. In: *American Journal of Roentgenology* 194.4, pp. 1095–1098.
Loughenbury, P. R., Wadhwani, S. and Soames, R. W. (Sept. 2006). 'The Posterior Longitudinal Ligament and Peridural (Epidural) Membrane'. In: *Clinical Anatomy* 19.6, pp. 487–492.
Love, T. W., Fagan, A. B. and Fraser, R. D. (1999). 'Degenerative Spondylolisthesis'. In: *THE JOURNAL OF BONE AND JOINT SURGERY* 81.4, p. 5.
Ltd, S. ((2020). *Affinity Designer*. Version 1.7.3.
Lurie, J. D. et al. (Apr. 2008a). 'Reliability of Magnetic Resonance Imaging Readings for Lumbar Disc Herniation in the Spine Patient Outcomes Research Trial (SPORT):' in: *Spine* 33.9, pp. 991–998.
Lurie, J. D. et al. (June 2008b). 'Reliability of Readings of Magnetic Resonance Imaging Features of Lumbar Spinal Stenosis:' in: *Spine* 33.14, pp. 1605–1610.
Luszczynska, A. et al. (Feb. 2011). 'Self-Efficacy as a Moderator of the Planning–Behaviour Relationship in Interventions Designed to Promote Physical Activity'. In: *Psychology & Health* 26.2, pp. 151–166.
Macedo, L. G. et al. (1st Dec. 2013a). 'Physical Therapy Interventions for Degenerative Lumbar Spinal Stenosis: A Systematic Review'. In: *Physical Therapy* 93.12, pp. 1646–1660.
Macedo, L. G., Wang, Y. and Battié, M. C. (May 2013b). 'The Sedimentation Sign for Differential Diagnosis of Lumbar Spinal Stenosis:' in: *Spine* 38.10, pp. 827–831.
Machado, G. C. et al. (1st Nov. 2016). 'Surgical Options for Lumbar Spinal Stenosis'. In: *Cochrane Database of Systematic Reviews*. Ed. by Cochrane Back and Neck Group.
Machiavelli, N. (1517). 'Discourses On Livy'. In: *Book I*.
Macnab, I. (Aug. 1950). 'Spondylolisthesis with an Intact Neural Arch; the so-Called Pseudo-Spondylolisthesis'. In: *The Journal of Bone and Joint Surgery. British Volume* 32-B.3, pp. 325–333. pmid: 14778851.
Mamisch, N. et al. (July 2012). 'Radiologic Criteria for the Diagnosis of Spinal Stenosis: Results of a Delphi Survey'. In: *Radiology* 264.1, pp. 174–179.
Manaka, M. et al. (Jan. 2003). 'Assessment of Lumbar Spinal Canal Stenosis by Magnetic Resonance Phlebography'. In: *Journal of Orthopaedic Science* 8.1, pp. 1–7.
Mannion, A. F. et al. (Oct. 2017). 'Dural Sac Cross-Sectional Area and Morphological Grade Show Significant Associations with Patient-Rated Outcome of Surgery for Lumbar Central Spinal Stenosis'. In: *European Spine Journal* 26.10, pp. 2552–2564.
Marawar, S. V. et al. (Dec. 2016). 'Comparison of Surgeon Rating of Severity of Stenosis Using Magnetic Resonance Imaging, Dural Cross-Sectional Area, and Functional Outcome Scores'. In: *World Neurosurgery* 96, pp. 165–170.
Mariconda, M. et al. (Feb. 2002). 'Unilateral Laminectomy for Bilateral Decompression of Lumbar Spinal Stenosis: A Prospective Comparative Study with Conservatively Treated Patients'. In: *Journal of Spinal Disorders & Techniques* 15.1, pp. 39–46. pmid: 11891449.
Marketos, S. G. and Skiadas, P. K. (Nov. 1999). 'Galen: A Pioneer of Spine Research:' in: *Spine* 24.22, p. 2358.
McCarron, R. F. et al. (Oct. 1987). 'The Inflammatory Effect of Nucleus Pulposus. A Possible Element in the Pathogenesis of Low-Back Pain'. In: *Spine* 12.8, pp. 760–764. pmid: 2961088.
McCulloch, J. (Feb. 1977). 'Chemonucleolysis'. In: *The Journal of Bone and Joint Surgery. British volume* 59-B.1, pp. 45–52.

Melzack, R. (1975). 'The McGill Pain Questionnaire: Major Properties and Scoring Methods'. In: *Pain* 1.3, pp. 277–299.
Menardi, G. and Torelli, N. (Jan. 2014). 'Training and Assessing Classification Rules with Imbalanced Data'. In: *Data Mining and Knowledge Discovery* 28.1, pp. 92–122.
Meyerding, H. W. (1932). 'Spondyloptosis'. In: *Surg Gynaecol Obstet.* 54, pp. 371–377.
Miao, J. et al. (June 2013). 'Segmental Spinal Canal Volume in Patients with Degenerative Spondylolisthesis'. In: *The Spine Journal* 13.6, pp. 706–712.
Mikhael, M. A. et al. (July 1981). 'Neuroradiological Evaluation of Lateral Recess Syndrome.' In: *Radiology* 140.1, pp. 97–107.
Min, J.-H., Jang, J.-S. and Lee, S.-H. (Jan. 2008). 'Clinical Significance of Redundant Nerve Roots of the Cauda Equina in Lumbar Spinal Stenosis'. In: *Clinical Neurology and Neurosurgery* 110.1, pp. 14–18.
Minamide, A., Yoshida, M. and Maio, K. (Sept. 2013). 'The Natural Clinical Course of Lumbar Spinal Stenosis: A Longitudinal Cohort Study over a Minimum of 10years'. In: *Journal of Orthopaedic Science* 18.5, pp. 693–698.
Mixter, W. J. and Barr, J. S. (2009). 'Rupture of the Intervertebral Disc with Involvement of the Spinal Canal'. In: *The New England Journal of Medicine*, p. 12.
Modic, M. T. et al. (Jan. 1988). 'Degenerative Disk Disease: Assessment of Changes in Vertebral Body Marrow with MR Imaging.' In: *Radiology* 166.1, pp. 193–199.
Mollayeva, T. et al. (Feb. 2016). 'The Pittsburgh Sleep Quality Index as a Screening Tool for Sleep Dysfunction in Clinical and Non-Clinical Samples: A Systematic Review and Meta-Analysis'. In: *Sleep Medicine Reviews* 25, pp. 52–73.
Moojen, W. A. et al. (Mar. 2018). 'Preoperative MRI in Patients With Intermittent Neurogenic Claudication: Relevance for Diagnosis and Prognosis'. In: *SPINE* 43.5, pp. 348–355.
Moon, E.-S. et al. (2005). 'Comparison of the Predictive Value of Myelography, Computed Tomography and MRI on the Treadmill Test in Lumbar Spinal Stenosis'. In: *Yonsei Medical Journal* 46.6, p. 806.
Moses, R. A. et al. (Feb. 2015). 'Is the Sedimentation Sign Associated With Spinal Stenosis Surgical Treatment Effect in SPORT?:' in: *Spine* 40.3, pp. 129–136.
Muto, M. et al. (2016a). 'Dynamic MR in Patients Affected by Neurogenical Claudication: Technique and Results from a Single-Center Experience'. In: *Neuroradiology* 58.8.
Muto, M. et al. (2016b). 'Neuroimaging of Spinal Instability'. In: *Magnetic Resonance Imaging Clinics of North America* 24.3, pp. 485–494.
Nachemson, A. and Morris, J. M. (1964). 'IN VIVO MEASUREMENTS OF INTRADISCAL PRESSURE. DISCOMETRY, A METHOD FOR THE DETERMINATION OF PRESSURE IN THE LOWER LUMBAR DISCS.' In: *The Journal of bone and joint surgery. American volume* 46, pp. 1077–92. PMID: 14193834.
Newell, A. M. et al. (1st Feb. 2012). 'The Modified Gait Efficacy Scale: Establishing the Psychometric Properties in Older Adults'. In: *Physical Therapy* 92.2, pp. 318–328.
Niosi, C. A. and Oxland, T. R. (Nov. 2004). 'Degenerative Mechanics of the Lumbar Spine'. In: *The Spine Journal* 4.6, S202–S208.
Nixon, J. E. (1987). 'Spinal Stenosis'.
Obuchowski, N. A. (2005). 'Fundamentals of Clinical Research for Radiologists'. In: p. 9.
OED-Online (2019a). *"disease, n."* Oxford University Press. URL: www.oed.com/view/Entry/54151.
— (2019b). *"stenosis, n"*. Oxford University Press. URL: www.oed.com/view/Entry/189805.
— (2019c). *Syndrome, n.* Oxford University Press. URL: http://www.oed.com/view/Entry/196454.
Ogilvie, D. et al. (Dec. 2008). 'The Harvest Plot: A Method for Synthesising Evidence about the Differential Effects of Interventions'. In: *BMC Medical Research Methodology* 8.1, p. 8.
Olmarker, K. et al. (1989). 'Effects of Experimental Graded Compression on Blood Flow in Spinal Nerve Roots. A Vital Microscopic Study on the Porcine Cauda Equina'. In: *Journal of Orthopaedic Research: Official Publication of the Orthopaedic Research Society* 7.6, pp. 817–823. pmid: 2795321.
Olmarker, K. and Rydevik, B. (June 1992). 'Single- Versus Double-Level Nerve Root Compression: An Experimental Study on the Porcine Cauda Equina With Analyses of Nerve Impulse Conduction Properties'. In: *Clinical Orthopaedics and Related Research* &NA;279, 35???39.
Omarker, K. and Myers, R. R. (Nov. 1998). 'Pathogenesis of Sciatic Pain: Role of Herniated Nucleus Pulposus and Deformation of Spinal Nerve Root and Dorsal Root Ganglion:' in: *Pain* 78.2, pp. 99–105.
Ooi, Y., Mita, F. and Satoh, Y. (June 1990). 'Myeloscopic Study on Lumbar Spinal Canal Stenosis with Special Reference to Intermittent Claudication'. In: *Spine* 15.6, pp. 544–549. pmid: 2402694.
Oppenheim, H. and Kause, F. (1909). 'Ueber Eintlemmung Bzw. Strangulation Der Cauda Equina'. In: *Deutsche Medizinische Wochenschrift* 35, pp. 697–700.
Organization, W. H. (2001). *International classification of functioning, disability and health: ICF*. Geneva: WHO.
Ozawa, H. et al. (June 2012). 'Dynamic Changes in the Dural Sac Cross-Sectional Area on Axial Loaded MR Imaging: Is There a Difference between Degenerative Spondylolisthesis and Spinal Stenosis?' In: *American Journal of Neuroradiology* 33.6, pp. 1191–1197.
Papp, T., Porter, R. W. and Aspden, R. M. (15th Dec. 1994). 'The Growth of the Lumbar Vertebral Canal'. In: *Spine* 19.24, pp. 2770–2773. pmid: 7899977.
— (May 1995). 'Trefoil Configuration and Developmental Stenosis of the Lumbar Vertebral Canal'. In: *The Journal of Bone and Joint Surgery. British Volume* 77.3, pp. 469–472. pmid: 7744939.
Park, C. H., Lee, S. H. and Jung, J. Y. (2011). 'Dural Sac Cross-Sectional Area Does Not Correlate with Efficacy of Percutaneous Adhesiolysis in Single Level Lumbar Spinal Stenosis'. In: *Pain Physician*, p. 6.
Park, C.-H. and Lee, S.-H. (Apr. 2014). 'Correlation Between Severity of Lumbar Spinal Stenosis and Lumbar Epidural Steroid Injection'. In: *Pain Medicine* 15.4, pp. 556–561.
Park, D. K. et al. (Feb. 2010). 'Does Multilevel Lumbar Stenosis Lead to Poorer Outcomes?: A Subanalysis of the Spine Patient Outcomes Research Trial (SPORT) Lumbar Stenosis Study'. In: *Spine* 35.4, pp. 439–446.
Parsons, S. et al. (Dec. 2006). 'Measuring Troublesomeness of Chronic Pain by Location'. In: *BMC Musculoskeletal Disorders* 7.1, p. 34.
Patrick, D. L. et al. (1st Sept. 1995). 'Assessing Health-Related Quality of Life in Patients with Sciatica'. In: *Spine* 20.17, 1899–1908, discussion 1909. pmid: 8560339.
Penning, L. (Feb. 1992). 'Functional Pathology of Lumbar Spinal Stenosis'. In: *Clinical Biomechanics* 7.1, pp. 3–17.
Penning, L. and Wilmink, J. T. (1981). 'Biomechanics of Lumbosacral Dural Sac. A Study of Flexion-Extension Myelography'. In: *Spine* 6.4, pp. 398–408. pmid: 7280829.
— (June 1987). 'Posture-Dependent Bilateral Compression of L4 or L5 Nerve Roots in Facet Hypertrophy. A Dynamic CT-Myelographic Study'. In: *Spine* 12.5, pp. 488–500. pmid: 3629399.
Pfirrmann, C. W. A. et al. (Sept. 2001). 'Magnetic Resonance Classification of Lumbar Intervertebral Disc Degeneration:' in: *Spine* 26.17, pp. 1873–1878.
Pfirrmann, C. W. A. et al. (Feb. 2004). 'MR Image–Based Grading of Lumbar Nerve Root Compromise Due to Disk Herniation: Reliability Study with Surgical Correlation'. In: *Radiology* 230.2, pp. 583–588.
Piperno, M. et al. (15th Sept. 1997). 'Phospholipase A2 Activity in Herniated Lumbar Discs. Clinical Correlations and Inhibition by Piroxicam'. In: *Spine* 22.18, pp. 2061–2065. pmid: 9322315.
Pope, M. H. and Panjabi, M. (Apr. 1985). 'Biomechanical Definitions of Spinal Instability'. In: *Spine* 10.3, pp. 255–256. pmid: 3992345.
Portal, A. (1804). *Cours d'anatomie Médicale: Ou Élémens de l'anatomie de l'homme...* Vol. 5. Baudouin.
Porter, R. W. (1st Sept. 1996). 'Spinal Stenosis and Neurogenic Claudication'. In: *Spine* 21.17, pp. 2046–2052. pmid: 8883210.
Porter, R. W., Hibbert, C. and Wellman, P. (1980). 'Backache and the Lumbar Spinal Canal'. In: *Spine* 5.2, pp. 99–105. pmid: 6446166.
Porter, R. W. and Pavitt, D. (Nov. 1987). 'The Vertebral Canal: I. Nutrition and Development, an Archaeological Study'. In: *Spine* 12.9, pp. 901–906. pmid: 3327173.
Porter, R. W. and Ward, D. (Jan. 1992). 'Cauda Equina Dysfunction. The Significance of Two-Level Pathology'. In: *Spine* 17.1, pp. 9–15. pmid: 1536018.
Porter, R. W., Wicks, M. and Ottewell, D. (Nov. 1978). 'Measurement of the Spinal Canal by Diagnostic Ultrasound'. In: *The*

Journal of Bone and Joint Surgery. British Volume 60-B.4, pp. 481–484. pmid: 711793.
Porter, R. W. (2001). 'Spinal Stenosis and Disorders of the Lumbar Spine'.
Porter, R., Hibbert, C. and Wicks, M. (Nov. 1978). 'The Spinal Canal in Symptomatic Lumbar Disc Lesions'. In: *The Journal of Bone and Joint Surgery. British volume* 60-B.4, pp. 485–487.
Postacchini, F., Ripani, M. and Carpano, S. (1983). 'Morphometry of the Lumbar Vertebrae. An Anatomic Study in Two Caucasoid Ethnic Groups'. In: *Clinical Orthopaedics and Related Research* 172, pp. 296–303. pmid: 6821998.
Prasad, B. C. M. et al. (2016). 'Clinical, Radiological, and Functional Evaluation of Surgical Treatment in Degenerative Lumbar Canal Stenosis'. In: *Neurology India* 64.4, p. 677.
Project, H. (2019). *Horos*. Version 3.3.5.
R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
Resnick, B. and Jenkins, L. S. (2000). 'Testing the Reliability and Validity of the Self-Efficacy for Exercise Scale'. In: *Nursing Research* 49.3, pp. 154–159. pmid: 10882320.
Richardson, W. F. and Carman, J. B. (1998). 'The Spinal Column, and Some General Remarks about the Bones Which Compose It'. In: *On the Fabric of the Human Body: A Translation of de Humani Corporis Fabrica Libri Septem*. San Fransisco CA: Normal Publishing, p. 136.
Roberts, H. C. et al. (1st Sept. 2012). 'Is Grip Strength Associated with Length of Stay in Hospitalised Older Patients Admitted for Rehabilitation? Findings from the Southampton Grip Strength Study'. In: *Age and Ageing* 41.5, pp. 641–646.
Roberts, S. et al. (15th Dec. 1995). 'Mechanoreceptors in Intervertebral Discs. Morphology, Distribution, and Neuropeptides'. In: *Spine* 20.24, pp. 2645–2651. pmid: 8747242.
Robin, X. et al. (2011). 'pROC: an open-source package for R and S+ to analyze and compare ROC curves'. In: *BMC Bioinformatics* 12, p. 77.
Roland, M. and Morris, R. (Mar. 1983). 'A Study of the Natural History of Back Pain. Part I: Development of a Reliable and Sensitive Measure of Disability in Low-Back Pain'. In: *Spine* 8.2, pp. 141–144. pmid: 6222486.
Ruscheweyh, R. et al. (Nov. 2009). 'Pain Sensitivity Can Be Assessed by Self-Rating: Development and Validation of the Pain Sensitivity Questionnaire:' in: *Pain* 146.1, pp. 65–74.
Rydevik, B., Brown, M. D. and Lundborg, G. (1984). 'Pathoanatomy and Pathophysiology of Nerve Root Compression'. In: *Spine* 9.1, pp. 7–15. pmid: 6372124.
Santiago, F. et al. (June 2001). 'Morphometry of the Lower Lumbar Vertebrae in Patients with and without Low Back Pain'. In: *European Spine Journal* 10.3, pp. 228–233.
Sarpyener, M. A. (July 1947). 'Spina Bifida Aperta and Congenital Stricture of the Spinal Canal'. In: *The Journal of Bone and Joint Surgery. American Volume* 29.3, pp. 817–821. pmid: 20253066.
Schellhas, K. P. et al. (1st Jan. 1996). 'Lumbar Disc High-Intensity Zone. Correlation of Magnetic Resonance Imaging and Discography'. In: *Spine* 21.1, pp. 79–86. pmid: 9122767.
Schellinger, D. et al. (June 1990). 'Disk Fragment Migration.' In: *Radiology* 175.3, pp. 831–836.
Schizas, C. et al. (Oct. 2010). 'Qualitative Grading of Severity of Lumbar Spinal Stenosis Based on the Morphology of the Dural Sac on Magnetic Resonance Images:' in: *Spine* 35.21, pp. 1919–1924.
Schizas, C. et al. (Aug. 2014). 'Secular Changes of Spinal Canal Dimensions in Western Switzerland: A Narrowing Epidemic?' In: *Spine* 39.17, pp. 1339–1344.
Schmid, M. R. et al. (Apr. 1999). 'Changes in Cross-Sectional Measurements of the Spinal Canal and Intervertebral Foramina as a Function of Body Position: In Vivo Studies on an Open-Configuration MR System.' In: *American Journal of Roentgenology* 172.4, pp. 1095–1102.
Schmidt, F. L. and Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications, Ltd.
Schnebel, B. et al. (Mar. 1989). 'Comparison of MRI to Contrast CT in the Diagnosis of Spinal Stenosis'. In: *Spine* 14.3, pp. 332–337. pmid: 2711248.
Schonstrom, N. S., Bolender, N. F. and Spengler, D. M. (Nov. 1985). 'The Pathomorphology of Spinal Stenosis as Seen on CT Scans of the Lumbar Spine'. In: *Spine* 10.9, pp. 806–811. pmid: 4089655.
Schönström, N. and Hansson, T. (Apr. 1988). 'Pressure Changes Following Constriction of the Cauda Equina. An Experimental Study in Situ'. In: *Spine* 13.4, pp. 385–388. pmid: 3406845.
Schönström, N. and Willén, J. (Jan. 2001). 'Imaging Lumbar Spinal Stenosis'. In: *Radiologic Clinics of North America* 39.1, pp. 31–53, v. pmid: 11221505.
Schönström, N. et al. (Sept. 1984). 'Pressure Changes within the Cauda Equina Following Constriction of the Dural Sac. An in Vitro Experimental Study'. In: *Spine* 9.6, pp. 604–607. pmid: 6495030.
Schrader, P. K. et al. (5th Aug. 1999). 'Histology of the Ligamentum Flavum in Patients with Degenerative Lumbar Spinal Stenosis'. In: *European Spine Journal* 8.4, pp. 323–328.
Serhan, H. A. et al. (Jan. 2007). 'Biomechanics of the Posterior Lumbar Articulating Elements'. In: *Neurosurgical Focus* 22.1, pp. 1–6.
Sharma, A., Parsons, M. and Pilgram, T. (Oct. 2011). 'Temporal Interactions of Degenerative Changes in Individual Components of the Lumbar Intervertebral Discs: A Sequential Magnetic Resonance Imaging Study in Patients Less Than 40 Years of Age'. In: *Spine* 36.21, pp. 1794–1800.
Shuval, K. et al. (Feb. 2014). 'Sedentary Behaviour and Physical Inactivity Assessment in Primary Care: The Rapid Assessment Disuse Index (RADI) Study'. In: *British Journal of Sports Medicine* 48.3, pp. 250–255.
Sigmundsson, F. G. et al. (Apr. 2011). 'Correlation between Disability and MRI Findings in Lumbar Spinal Stenosis: A Prospective Study of 109 Patients Operated on by Decompression'. In: *Acta Orthopaedica* 82.2, pp. 204–210.
— (Oct. 2012). 'Prognostic Factors in Lumbar Spinal Stenosis Surgery: A Prospective Study of Imaging- and Patient-Related Factors in 109 Patients Who Were Operated on by Decompression'. In: *Acta Orthopaedica* 83.5, pp. 536–542.
Singh, K. et al. (Nov. 2005). 'Congenital Lumbar Spinal Stenosis: A Prospective, Control-Matched, Cohort Radiographic Analysis'. In: *The Spine Journal* 5.6, pp. 615–622.
Sirvanci, M. et al. (May 2008). 'Degenerative Lumbar Spinal Stenosis: Correlation with Oswestry Disability Index and MR Imaging'. In: *European Spine Journal* 17.5, pp. 679–685.
Smyth, M. J. and Wright, V. (Dec. 1958). 'Sciatica and the Intervertebral Disc; an Experimental Study'. In: *The Journal of Bone and Joint Surgery. American Volume* 40-A.6, pp. 1401–1418. pmid: 13610969.
Splendiani, A. et al. (Mar. 2014). 'Occult Neural Foraminal Stenosis Caused by Association between Disc Degeneration and Facet Joint Osteoarthritis: Demonstration with Dedicated Upright MRI System'. In: *La radiologia medica* 119.3, pp. 164–174.
Steurer, J. et al. (Dec. 2011). 'Quantitative Radiologic Criteria for the Diagnosis of Lumbar Spinal Stenosis: A Systematic Literature Review'. In: *BMC Musculoskeletal Disorders* 12.1, p. 175.
Strojnik, T. (2001). 'Measurement of the Lateral Recess Angle as a Possible Alternative for Evaluation of the Lateral Recess Stenosis on a CT Scan'. In: *Wiener Klinische Wochenschrift* 113 Suppl 3, pp. 53–58. pmid: 15503622.
Stucki, G. et al. (1st Apr. 1996). 'Measurement Properties of a Self-Administered Outcome Measure in Lumbar Spinal Stenosis'. In: *Spine* 21.7, pp. 796–803. pmid: 8779009.
Suh, S. W. et al. (Apr. 2005). 'Origin of Lumbar Spinal Roots and Their Relationship to Intervertebral Discs: A CADAVER AND RADIOLOGICAL STUDY'. In: *The Journal of Bone and Joint Surgery. British volume* 87-B.4, pp. 518–522.
Tait, R. C., Chibnall, J. T. and Krause, S. (Feb. 1990). 'The Pain Disability Index: Psychometric Properties:' in: *Pain* 40.2, pp. 171–182.
Takahashi, K. et al. (15th Dec. 1995). 'Changes in Epidural Pressure during Walking in Patients with Lumbar Spinal Stenosis'. In: *Spine* 20.24, pp. 2746–2749. pmid: 8747254.
Teraguchi, M. et al. (July 2017). 'Progression, Incidence, and Risk Factors for Intervertebral Disc Degeneration in a Longitudinal Population-Based Cohort: The Wakayama Spine Study'. In: *Osteoarthritis and Cartilage* 25.7, pp. 1122–1131.
Tew, G. et al. (May 2013). 'Feasibility and Validity of Self-Reported Walking Capacity in Patients with Intermittent Claudication'. In: *Journal of Vascular Surgery* 57.5, pp. 1227–1234.
Therneau, T. and Atkinson, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Thornicroft, G., Beecham, J. and Knapp, M. (2001). 'Costing Psychiatric Interventions'. In: *Measuring Mental Health Needs London: Gaskell*, pp. 220–4.

Tomkins-Lane, C. C. et al. (Apr. 2012). 'Predictors of Walking Performance and Walking Capacity in People With Lumbar Spinal Stenosis, Low Back Pain, and Asymptomatic Controls'. In: *Archives of Physical Medicine and Rehabilitation* 93.4, pp. 647–653.

Tomkins-Lane, C. C. et al. (Nov. 2013). 'Nerve Root Sedimentation Sign for the Diagnosis of Lumbar Spinal Stenosis: Reliability, Sensitivity, and Specificity'. In: *Spine* 38.24, E1554–E1560.

Ullrich, C. G. et al. (Jan. 1980). 'Quantitative Assessment of the Lumbar Spinal Canal by Computed Tomography.' In: *Radiology* 134.1, pp. 137–143.

Ulmer, J. L. et al. (Aug. 1994). 'Distinction between Degenerative and Isthmic Spondylolisthesis on Sagittal MR Images: Importance of Increased Anteroposterior Diameter of the Spinal Canal ("wide Canal Sign").' In: *American Journal of Roentgenology* 163.2, pp. 411–416.

Ulmer, J. L. et al. (1995). 'Lumbar Spondylolysis without Spondylolisthesis: Recognition of Isolated Posterior Element Subluxation on Sagittal MR'. In: p. 6.

Urquhart, D. M. et al. (Dec. 2015). 'Could Low Grade Bacterial Infection Contribute to Low Back Pain? A Systematic Review'. In: *BMC Medicine* 13.1, p. 13.

Ursu, T. R., Porter, R. W. and Navaratnam, V. (1st Dec. 1996). 'Development of the Lumbar and Sacral Vertebral Canal in Utero'. In: *Spine* 21.23, pp. 2705–2708. pmid: 8979314.

Ustün, T. B. et al. (2003). 'The International Classification of Functioning, Disability and Health: A New Tool for Understanding Disability and Health'. In: *Disability and Rehabilitation* 25.11-12, pp. 565–571. pmid: 12959329.

Van Gelderen, C. (1948). 'Ein Orthotisches (Lordotisches) Kaudasyndrom'. In: *Acta Psychiat* 23.57.

Vandenabeele, F., Creemers, J. and Lambrichts, I. (1996). 'Ultrastructure of the Human Spinal Arachnoid Mater and Dura Mater'. In: p. 14.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with s*. 4th ed. New York: Springer.

Verbiest, H. (May 1954). 'A RADICULAR SYNDROME FROM DEVELOPMENTAL NARROWING OF THE LUMBAR VERTEBRAL CANAL'. In: *The Journal of Bone and Joint Surgery. British volume* 36-B.2, pp. 230–237.

— (1976). 'Fallacies of the Present Definition, Nomenclature, and Classification of the Stenoses of the Lumbar Vertebral Canal'. In: *Spine* 1.4, pp. 217–225.

— (1979). 'The Significance and Principles of Computerized Axial Tomography in Idiopathic Developmental Stenosis of the Bony Lumbar Vertebral Canal'. In: *Spine* 4.4, pp. 369–378. pmid: 483044.

Verhoof, O. J. et al. (Feb. 2008). 'High Failure Rate of the Interspinous Distraction Device (X-Stop) for the Treatment of Lumbar Spinal Stenosis Caused by Degenerative Spondylolisthesis'. In: *European Spine Journal* 17.2, pp. 188–192.

Veritas Health Innovation Melbourne, A. (2017). *Covidence systematic review software*. Available at http://www.covidence.org.

Verleysen, M. and François, D. (2005). 'The Curse of Dimensionality in Data Mining and Time Series Prediction'. In: *Computational Intelligence and Bioinspired Systems*. Ed. by J. Cabestany, A. Prieto and F. Sandoval. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 758–770.

Viale, G. L. (June 2004). 'THE SPINAL CORD AND ITS ROOTS ACCORDING TO GALEN'. In: *Neurosurgery* 54.6, pp. 1490–1496.

Viechtbauer, W. (2010). 'Conducting meta-analyses in R with the metafor package'. In: *Journal of Statistical Software* 36.3, pp. 1–48.

Waddell, G. et al. (Feb. 1993). 'A Fear-Avoidance Beliefs Questionnaire (FABQ) and the Role of Fear-Avoidance Beliefs in Chronic Low Back Pain and Disability'. In: *Pain* 52.2, pp. 157–168. pmid: 8455963.

Wang, F. et al. (Mar. 2016). 'Aging and Age Related Stresses: A Senescence Mechanism of Intervertebral Disc Degeneration'. In: *Osteoarthritis and Cartilage* 24.3, pp. 398–408.

Wang, J. and Yang, X. (Aug. 2009). 'Age-Related Changes in the Orientation of Lumbar Facet Joints:' in: *Spine* 34.17, E596–E598.

Ward, L. et al. (June 2019). 'Development and Delivery of the BOOST (Better Outcomes for Older Adults with Spinal Trouble) Intervention for Older Adults with Neurogenic Claudication'. In: *Physiotherapy* 105.2, pp. 262–274.

Ware, J. E. and Sherbourne, C. D. (1992). 'The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection'. In: *Medical Care* 30.6, pp. 473–483. JSTOR: 3765916.

Weber, C. et al. (Jan. 2016). 'Is There an Association Between Radiological Severity of Lumbar Spinal Stenosis and Disability, Pain, or Surgical Outcome?: A Multicenter Observational Study'. In: *SPINE* 41.2, E78–E83.

Weiner, B. K., Patel, N. M. and Walker, M. A. (Dec. 2007). 'Outcomes of Decompression for Lumbar Spinal Canal Stenosis Based upon Preoperative Radiographic Severity'. In: *Journal of Orthopaedic Surgery and Research* 2.1, p. 3.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

— (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Wildermuth, S. et al. (May 1998). 'Lumbar Spine: Quantitative and Qualitative Assessment of Positional (Upright Flexion and Extension) MR Imaging and Myelography.' In: *Radiology* 207.2, pp. 391–398.

Willén, J. and Danielson, B. (Dec. 2001). 'The Diagnostic Effect From Axial Loading of the Lumbar Spine During Computed Tomography and Magnetic Resonance Imaging in Patients With Degenerative Disorders:' in: *Spine* 26.23, pp. 2607–2614.

Williamson, E. et al. (Oct. 2018). 'Better Outcomes for Older People with Spinal Trouble (BOOST) Trial: A Randomised Controlled Trial of a Combined Physical and Psychological Intervention for Older Adults with Neurogenic Claudication, a Protocol'. In: *BMJ Open* 8.10, e022205.

Wilmink, J. T., Korte, J. H. and Penning, L. (Dec. 1988). 'Dimensions of the Spinal Canal in Individuals Symptomatic and Non-Symptomatic for Sciatica: A CT Study'. In: *Neuroradiology* 30.6, pp. 547–550.

Wilmink, J. T. (2010a). 'Imaging Techniques for the Lumbar Spine: Conventional Radiology, Computed Tomography; Magnetic Resonance Imaging'. In: *Lumbar Spinal Imaging in Radicular Pain and Related Conditions*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 9–30.

— (2010b). 'Normal Anatomy'. In: *Lumbar Spinal Imaging in Radicular Pain and Related Conditions*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 31–57.

— (2010c). 'Pathologic Anatomy and Mechanisms of Nerve Root Compression'. In: *Lumbar Spinal Imaging in Radicular Pain and Related Conditions*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 59–113.

— (2010d). 'The Nature of Radicular Pain and Related Conditions'. In: *Lumbar Spinal Imaging in Radicular Pain and Related Conditions*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–7.

Wiltse, L. L. (Apr. 1962). 'The Etiology of Spondylolisthesis'. In: *The Journal of Bone and Joint Surgery. American Volume* 44-A, pp. 539–560. pmid: 14040287.

Winston, K., Rumbaugh, C. and Colucci, V. (1984). 'The Vertebral Canals in Lumbar Disc Disease'. In: *Spine* 9.4, pp. 414–417. pmid: 6474255.

Witten, I. H., ed. (2017). *Data Mining: Practical Machine Learning Tools and Techniques*. Fourth Edition. Amsterdam: Elsevier. 621 pp.

Wiyanad, A. et al. (Dec. 2018). 'Is the Occiput-Wall Distance Valid and Reliable to Determine the Presence of Thoracic Hyperkyphosis?' In: *Musculoskeletal Science & Practice* 38, pp. 63–68. pmid: 30278368.

Wong, A. Y., Karppinen, J. and Samartzis, D. (Dec. 2017). 'Low Back Pain in Older Adults: Risk Factors, Management Options and Future Directions'. In: *Scoliosis and Spinal Disorders* 12.1, p. 14.

Wright, M. N. and Ziegler, A. (2017). 'Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R'. In: *Journal of Statistical Software* 77.1.

Yamashita, K. et al. (Nov. 2003). 'Correlation of Patient Satisfaction with Symptom Severity and Walking Ability after Surgical Treatment for Degenerative Lumbar Spinal Stenosis:' in: *Spine* 28.21, pp. 2477–2481.

Yoshida, M. et al. (Nov. 1992). 'Hypertrophied Ligamentum Flavum in Lumbar Spinal Canal Stenosis. Pathogenesis and

Morphologic and Immunohistochemical Observation'. In: *Spine* 17.11, pp. 1353–1360. pmid: 1462211.

Zaina, F. et al. (29th Jan. 2016). 'Surgical versus Non-Surgical Treatment for Lumbar Spinal Stenosis'. In: *Cochrane Database of Systematic Reviews*. Ed. by Cochrane Back and Neck Group.

Zanoli, G., Strömqvist, B. and Jönsson, B. (Nov. 2001). 'Visual Analog Scales for Interpretation of Back and Leg Pain Intensity in Patients Operated for Degenerative Lumbar Spine Disorders:' in: *Spine* 26.21, pp. 2375–2380.

Zeifang, F. et al. (Dec. 2008). 'Gait Analysis Does Not Correlate with Clinical and MR Imaging Parameters in Patients with Symptomatic Lumbar Spinal Stenosis'. In: *BMC Musculoskeletal Disorders* 9.1, p. 89.

Zhang, L. et al. (Oct. 2017). 'The Nerve Root Sedimentation Sign for Differential Diagnosis of Lumbar Spinal Stenosis: A Retrospective, Consecutive Cohort Study'. In: *European Spine Journal* 26.10, pp. 2512–2519.

Zhang, Q., Bhalerao, A. and Hutchinson, C. (Aug. 2017). 'Deformable Appearance Pyramids for Anatomy Representation, Landmark Detection and Pathology Classification'. In: *International Journal of Computer Assisted Radiology and Surgery* 12.8, pp. 1271–1280.

Zhao, C.-Q. et al. (Oct. 2007). 'The Cell Biology of Intervertebral Disc Aging and Degeneration'. In: *Ageing Research Reviews* 6.3, pp. 247–261.

Colophon

This thesis was written in GNU Emacs and typeset using LaTeX. The main font is *Valkyrie* with use of *Heliotrope* for headings and *Triplicate* for source-code blocks. These fonts were designed by Mathew Butterick, whose practicaltypography.com was an invaluable resource during the final preparation of the thesis. All typesetting mistakes, of which there are no doubt many, remain my own.

Plots were produced with a mixture of *ggplot2* (Wickham 2009) using the clean theme from *ggthmes* (Arnold 2019). ROC plots were produced using the *pROC* package (Robin et al. 2011). Other graphics were produced using *Affinity Designer* ® 2020, or, where not my own, are used with kind permission from the copyright holder as detailed in the relevant figure legends.

References were managed using Zotero and BibLaTeX.